FACULTADE DE FILOLOXÍA

Grao en Lingua e Literatura Españolas

Traballo de Fin de Grao

Estrategias de atribución automática de autoría aplicadas al teatro de Tirso de Molina

Miguel Cavadas Docampo
Autor

Pablo Gamallo Otero
Director

Santiago de Compostela

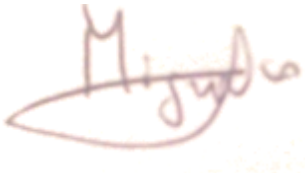
Curso 2018-2019

FACULTADE DE FILOLOXÍA

Grao en Lingua e Literatura Españolas

Traballo de Fin de Grao

Estrategias de atribución automática de autoría aplicadas al teatro de Tirso de Molina



GAMALLO OTERO PABLO - 36099559R
Firmado digitalmente por GAMALLO OTERO PABLO - 36099559R
Fecha: 2019.06.20 13:57:42 +02'00'

Miguel Cavadas Docampo
Autor

Pablo Gamallo Otero
Director

Santiago de Compostela

Curso 2018-2019

ÍNDICE

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA FACULTADE DE FILOLOXIA

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE FILOLOXÍA
- 6 NOV. 2018
ENTRADA
Nº

CUBRIR ESTE FORMULARIO ELECTRONICAMENTE

Formulario de delimitación de título e resumo
Traballo de Fin de Grao curso 2018/2019

APELIDOS E NOME: CAVADAS DOCAMPO, MIGUEL

GRAO EN: LINGUA E LITERATURA ESPAÑOLAS

(NO CASO DE MODERNAS) MENCIÓN EN:

TITOR/A: PABLO GAMALLO OTERO

LIÑA TEMÁTICA ASIGNADA: LINGÜÍSTICA COMPUTACIONAL

SOLICITO a aprobación do seguinte título e resumo:

Título:

Estrategias de atribución automática de autoría aplicadas al teatro de Tirso de Molina

Resumo:

La abundante producción dramática de Tirso de Molina (1579-1648) cuenta con varios casos de autoría dudosa que tradicionalmente se han venido atribuyendo al mercedario madrileño sin una base documental sólida. Entre ellos destaca la célebre obra *El burlador de Sevilla*, fundadora del mito de Don Juan, que un cierto sector de la crítica atribuye al dramaturgo Andrés de Claramonte (c. 1560-1626) y que además presenta otros problemas ecdóticos añadidos a raíz de la aparición de una versión alternativa del texto publicada bajo el título de *Tan largo me lo fiáis*. Ambos fenómenos evidencian el intrincado proceso de transmisión textual de las comedias del Siglo de Oro, germen de multitud de discusiones entre la crítica.

El propósito de este trabajo es configurar un corpus de comedias con el fin de realizar análisis comparativos de estilo a través de diferentes herramientas informáticas basadas en técnicas cuantitativas y estadísticas, de manera que se puedan extraer resultados numéricos que, interpretados en su contexto, sirvan como indicios para reforzar alguna de las múltiples posturas defendidas en este controvertido debate.

Santiago de Compostela, a 4 de novembro de 2018.

Sinatura do/a interesado/a	Visto e prace (sinatura do/a titor/a)	Aprobado pola Comisión de Títulos de Grao con data
	GAMALLO OTERO PABLO - 36099559R Firmado digitalmente por GAMALLO OTERO PABLO - 36099559R Fecha: 2018.11.04 22:59:59 +01'00'	16 NOV. 2018 Selo da Facultade de Filoloxía

SRA. DECANA DA FACULTADE DE FILOLOXÍA (Presidenta da Comisión de Títulos de Grao)

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA FACULTADE DE FILOLOXÍA

1. INTRODUCCIÓN

Uno de los principales problemas a los que la crítica textual debe enfrentarse a la hora de editar un texto antiguo es la correcta determinación de su autor. Los textos teatrales, por su propia concepción ligada a la escena y el intrincado proceso de transmisión textual que en su mayoría padecieron, exhiben una mayor dificultad que en muchos casos lleva a los investigadores por caminos interminables y baldíos sin una expectativa factible de encontrar una solución definitiva al dilema. La producción de Tirso de Molina (1579-1648), seudónimo con el que se conoce a fray Gabriel Téllez, es un claro ejemplo de este laberinto ecdótico plagado de recovecos y ramificaciones. Son muchos los críticos que han dedicado estudios a esclarecer los varios problemas de autoría que rodean a la obra del dramaturgo, pero en general los resultados no son, en absoluto, concluyentes. El principal obstáculo que frena la llegada de una resolución quizá sea que, en muchas ocasiones, el método empleado no posee un grado de discriminación lo suficientemente determinante como para proponer una atribución consistente. Por esto, una primera motivación de este trabajo es la formulación de una nueva forma de aproximarse a este debate teórico con herramientas que provienen de la lingüística computacional.

Prueba de que el debate tirsiano está más vivo que nunca es que el investigador de la Universidad de Salamanca Alejandro García-Reidy acaba de publicar en abril de 2019 un artículo en el que demuestra la autoría de Lope de la obra *Siempre ayuda la verdad*, tradicionalmente atribuida a Tirso, empleando entre otros métodos el software *Stylo* de R, al que también se ha recurrido para la realización de este trabajo. En su trabajo, García-Reidy habla de un estudio aún no publicado («¿Arbitraria y conjetural?: la estilometría en los estudios de autoría del teatro del Siglo de Oro») de Patricia Marín Cepeda que, según afirma, también aplica *Stylo* a un corpus amplio de obras dramáticas auriseculares. Teniendo en cuenta estas circunstancias, parece el momento idóneo para plantear una investigación que impulse el interés por la cuestión tirsiana y que siga un patrón alternativo.

Partiendo de la neutralidad (condición necesaria que, sin embargo, no ha sido siempre respetada en este debate), el objetivo de este trabajo no puede ser demostrar que X o Y obra no es de Tirso, ni menos arriesgar una atribución novedosa. Los propósitos deben limitarse a la extracción de conclusiones pertinentes que sirvan para reforzar

alguna de las posturas críticas del debate. No menos importante es la finalidad de demostrar la utilidad de las estrategias seleccionadas y fomentar su aplicación en investigaciones humanísticas o, en otras palabras, contribuir al desarrollo de las humanidades digitales, disciplina incipiente que todavía tiene mucho por decir en cuanto a las metodologías empleadas en la investigación literaria. Más allá de los estudios de atribución de autoría, muchas otras ramas de los estudios literarios se beneficiarían notablemente del recurso a medios informáticos. Se sitúa así este trabajo en los inicios de una nueva etapa académica que instituye la interdisciplinaridad como premisa esencial para su desarrollo.

La principal contribución de este estudio es, por consiguiente, la reorientación del debate tirsiano hacia el ámbito de las humanidades digitales. A pesar de manejar diversas estrategias que se basan en métodos numéricos y estadísticos que difieren entre sí considerablemente, los resultados obtenidos alcanzan un grado de consenso bastante elevado, de manera que permiten extraer conclusiones coherentes sobre las hipótesis que la crítica tradicional lleva siglos proponiendo. Cabe destacar, como adelanto, que dichas hipótesis suelen carecer de una base documental sólida y que varias de ellas quedan prácticamente desechadas a la vista de los resultados, que a su vez fuerzan a poner el foco sobre figuras más veladas, como lo es el dramaturgo murciano Andrés de Claramonte, muy posiblemente el verdadero autor del archiconocido drama *El burlador de Sevilla*.

Antes de llegar a esos resultados, se expone el marco teórico de los NTAAS (*non-traditional authorship attribution studies*) en el apartado 2, con el fin de ilustrar el estado de la disciplina y los trabajos relacionados gracias a lo que es posible abordar este con unas expectativas esperanzadoras. A continuación, en el apartado 3 se describe el marco metodológico de las estrategias empleadas y se verifica su utilidad en un corpus de entrenamiento. En el apartado 4, la parte nuclear del trabajo, se aplican las estrategias al teatro de Tirso de Molina, previa exposición del estado de la cuestión del debate y del proceso de configuración del corpus de prueba. Tras una discusión conjunta de todos los resultados obtenidos cierra finalmente el trabajo una sección de conclusiones que concentra las consecuencias que se han derivado del estudio y traza un viable plan de futuro para atajar el problema y acelerar la llegada de su resolución definitiva.

2. TRABAJO RELACIONADO: LOS *NON-TRADITIONAL AUTHORSHIP ATTRIBUTION STUDIES*

Los problemas de autoría han sido una constante a lo largo de la historia de la literatura. Tanto por prohibiciones legales como por procesos de transmisión textual enrevesados, por cuestiones de apropiaciones indebidas de textos ajenos o por colaboraciones entre varios autores, los casos de atribución de autoría dudosa constituyen un porcentaje nada desdeñable del conjunto de obras canónicas de la literatura mundial. Los investigadores que se han adentrado en el escabroso terreno de los estudios de atribución de autoría han desarrollado y perfeccionado una serie de métodos basados en la comparación estadística de diversos marcadores de estilo que supuestamente arrojan luz sobre el particular idiolecto de cada escritor. Entre esos marcadores se encuentra la métrica, la longitud de palabras u oraciones, la puntuación, los índices léxicos, los errores, las clases de palabras, las colocaciones y otros patrones morfosintácticos. Sin embargo, el análisis de esos elementos no ha conducido, por lo general, a casi ninguna conclusión taxativa, revelando su insuficiencia metodológica. Afortunadamente, los avances informáticos han favorecido la eclosión de campos interdisciplinares que funcionan como zonas de convergencia entre distintos ámbitos cuya cooperación puede resultar en un avance para ambos. De esta manera, las humanidades digitales, absorbiendo métodos y teorías provenientes de la lingüística computacional, se han erigido en toda una nueva forma de investigación literaria que aún no ha expresado todo su potencial. Su aplicación a los problemas de autoría ha dado lugar a la disciplina conocida como estudios de atribución de autoría no tradicionales (*non-traditional authorship attribution studies*, NTAAS a partir de ahora).

Los NTAAS nacieron bajo la expectativa de superar a los estudios tradicionales; sin embargo, a pesar de sus aproximadamente cincuenta años de trayectoria, no han conseguido establecer una cimentación firme que permitiese iniciar una línea de progresión ni eludir todo tipo de cuestionamientos y críticas. Según Joseph Rudman, «non-traditional authorship attribution studies have had enough time to pass through any "shake-down" phase and enter one marked by solid, scientific, and steadily progressing studies. But, after over 30 years and 300 publications, they have not» (1998: 351). La situación en estos últimos veinte años no ha variado mucho. Entre los problemas de los que adolece la especialidad, Rudman cita: «studies governed by expediency; a lack of competent research; flawed statistical techniques; corrupted

primary data; lack of expertise in allied fields; a dilettantish approach; inadequate treatment of errors» (351). Estas flaquezas se deben, en general, al enfoque científico con el que requieren ser abordados los NTAAS, a diferencia de otras investigaciones humanísticas. Sirva a modo de divisa la siguiente afirmación de Rudman: «Every non-traditional authorship attribution study is an experiment, a "scientific" experiment» (2016: 310). Como experimento científico cada estudio debe cumplir a rajatabla dos principios: replicabilidad y un plan experimental. Todos los experimentos cuya realización refleja este trabajo son fácilmente replicables y siguen un meticuloso diseño que se explicitará con detalle más adelante.

Las soluciones que propone Rudman para esta deficiencia son «construct a correct and complete experimental design; educate the practitioners; study style in its totality; identify and educate the gatekeepers; develop a complete theoretical framework; form an association of practitioners» (1998: 351). Solo siguiendo esas pautas se podrá reforzar el alcance de una disciplina que en su propia concepción contiene el componente clave que nunca podrá ser alcanzado por un estudio convencional: la automatización. Jack Grieve recuerda que «only quantitative techniques may be empirically evaluated and mechanically applied» (2005: 2), de ahí la necesidad de desarrollarlas, y aclara las ventajas de la automatización en el futuro: «At this time, automation primarily allows for investigators to rigorously test and fine-tune their techniques, but as these techniques improve, automation will also allow for unsupervised attribution applications» (117). Entre esas aplicaciones no solo se encuentra la verificación de las atribuciones, sino también «plagiarism detection, autor profiling or characterization, detection of stylistic inconsistencies in collaborative writing» (Stamatos, 2009: 539) o casos relativos a la lingüística forense.

Los NTAAS se han desarrollado fundamentalmente en inglés, con multitud de estudios que pretenden definir al fin el verdadero canon de textos shakespearianos. En español han sido estudiados recientemente, con mayor o menor éxito, los dos grandes enigmas autorales de la literatura española: la identidad de Avellaneda (Blasco, 2016) y la del autor del *Lazarillo* (La Rosa, 2016). En 2017 surgió el proyecto EstilometríaTSO (<http://estilometriatso.com/>), sobre teatro aurisecular, de mano de Álvaro Cuéllar González y Germán Vega García-Luengos, por lo que la investigación todavía está en una fase muy incipiente. No es en absoluto alentador el panorama al que se enfrenta el investigador que se atreva a afrontar este reto, pues como denuncia José Calvo Tello,

«una de las principales dificultades para trabajar en diferentes métodos de Humanidades Digitales con textos en español es precisamente la falta de textos literarios disponibles en formatos aceptables» (2016: 146), lo que fuerza a cada investigador a «encargarse de conseguir un texto digno a partir de las obras troceadas en diferentes páginas HTML en formato hoy en día obsoleto» (147). La recolección de textos para este trabajo, detallada en el apartado 4.2, es un claro ejemplo de ese camino repleto de obstáculos que dificulta enormemente la concentración en la parte experimental, que debería ser la nuclear en cualquier estudio de estas características.

La asunción básica de los NTAAS podría ser resumida en esta frase de Grieve: «the author of a text can be selected from a set of possible authors by comparing the values of textual measurements in that text to their corresponding values in each author's writing sample» (2005: 1). Esas medidas textuales se apoyan en diversas fórmulas matemáticas y sirven a propósitos diferentes. La selección de una u otra dependerá del corpus disponible y el objetivo del estudio, pero como concluye Grieve en su tesina, «the best approach to quantitative authorship attribution is one that is based on the values of as many textual measurements as posible» (118), ya que, en general, pequeñas variaciones en las configuraciones pueden producir cambios muy grandes en los resultados. Por esta razón, en este trabajo se emplean cinco estrategias diferentes y se cotejan conjuntamente sus resultados en el apartado 4.4.

Conviene tener en cuenta que los NTAAS siempre precisan de un pormenorizado estudio convencional previo que esclarezca cuáles deben ser los parámetros empleados: «a traditional authorship study looking at all of the external and traditional internal evidence must be completed before a non-traditional study is undertaken» (Rudman, 1998: 359) y que, por supuesto, conllevan las limitaciones intrínsecas a los estudios de atribución de autoría independientemente del método seguido. Por ejemplo, la inexistencia de textos a la hora de la comparación: «if […] the *Lazarillo* was the only work written by his author, any method, computational or not, based on the comparison of styles […] turns out to be useless» (La Rosa, 2016: 52), que es la aplicación práctica del presupuesto teórico «it is important to unmask the anonymous of a work if the writer is in fact an important author» (52). Por todo esto, no se deben despreciar las conclusiones derivadas de los NTAAS ni tampoco esperar milagros de ellos, sino considerar sus resultados y cotejarlos con los de estudios más tradicionales.

Los NTAAS pueden ser cualitativos, si identifican y describen los elementos característicos de un autor, o cuantitativos si miden de alguna forma esos elementos, aunque en la práctica lo común es un abordaje mixto (Almeida, 2014: 14) que siga el siguiente esquema:

> In every authorship identification problem, there is a set of candidate authors, a set of text samples of known authorship covering all the candidate authors (training corpus), and a set of text samples of unknown authorship (test corpus), each one of them should be attributed to a candidate author (Stamatos, 2009: 549).

Otra clasificación operativa es la de enfoque por modelo / enfoque por caso, que supone «the most basic property of the attribution methods since it largely determines the philosophy of each method» (549). El enfoque por modelo (*profile-based approach*) no tiene en cuenta las diferencias entre textos escritos por el mismo autor, empleando para cada autor un solo archivo enorme construido encadenando todos los textos disponibles para ese autor. En cambio, el enfoque por caso (*instance-based approach*) «requires multiple training text samples per author in order to develop an accurate attribution model» (549). En este trabajo se emplea tanto el enfoque por modelo como el enfoque por caso con el fin de contrastar los resultados obtenidos por dos aproximaciones basadas en concepciones antitéticas.

Los NTAAS se basan en técnicas estilométricas, esto es, técnicas estadísticas cuyo propósito es cuantificar el estilo, al que se le otorga de este modo poder discriminatorio. En palabras de Dayane Celestino de Almeida, «a idea de que é possível saber se alguém é o autor de um texto ou um conjunto de textos baseia-se no pressuposto de que o uso linguístico torna-se, ao longo do tempo, um hábito» (2014: 152) y por esta razón se convierte en una huella propia que distingue a todos los hablantes de una misma lengua. Almeida recuerda que «o estilo linguístico é dado pela recorrência de uma combinação de traços linguísticos, uma "constelação de variáveis", e não pelo estabelecimento de um único traço isoladamente» (154), por lo que la identificación de un autor siempre se basará en el cotejo conjunto de una amplia serie de rasgos. Aun así, existen detractores de esta metodología que aducen como argumentos su escaso cientifismo y su ignorancia de la variación lingüística intrahablante (2014: 20). Sin embargo, la estilometría no pretende alcanzar el estatus de ciencia, sino simplemente ofrecer «a mathematisation of stylistics – a new way of discriminating between forms of language behaviour that is of great potential value but not as yet a way of accounting for them» (Love, 2002: 161), un intento de «descobrir marcadores

quantificáveis de autoría […] frequentemente com auxílio computacional» (Almeida, 2014: 10).

La pregunta ahora es: ¿dónde radica exactamente ese carácter medible del estilo? Según Efstathios Stamatos, «the most common words (articles, prepositions, pronouns, etc.) are found to be among the best features to discriminate between authors» (2009: 542), es decir, las palabras estructurales, que no aportan información semántica, de lo que se deduce:

> Style-based text classification using lexical features require much lower dimensionality in comparison to topic-based text classification. In other words, much less words are sufficient to perform authorship attribution (a few hundred words) in comparison to a thematic text categorization task (several thousand words). More importantly, function words are used in a largely unconscious manner by the authors and they are topic-independent. Thus, they are able to capture pure stylistic choices of the authors across different topics (542).

Además, las palabras estructurales no solo actúan como un indicador léxico, sino que también contienen un alto grado de información sintáctica, ya que participan en una elevada cantidad de estructuras sintácticas. Por tanto, parece que, desde este punto de vista, extraer las palabras más frecuentes de un texto será el primer paso para emprender un cotejo estilométrico. Como recuerda Stamatos, «the most important criterion for selecting features in authorship attribution tasks is their frequency. In general, the more frequent a feature, the more stylistic variation it captures» (548). Conviene matizar que no se trata simplemente de elaborar una lista de términos, sino que es su distribución en los textos lo que constituye la huella del autor; no solo qué palabras usa más sino cuánto las usa. Con todo, la frecuencia no es el único factor determinante; también entra en juego la inestabilidad:

> Given a number of variations of the same text, all with the same meaning, the features that remain practically unchanged in all texts are considered stable. In other words, stability may be viewed as the availability of 'synonyms' for certain language characteristics. For example, words like 'and' and 'the' are very stable since there are no alternatives for them. On the other hand, words like 'benefit' or 'over' are relatively unstable since they can be replaced by 'gain' and 'above', respectively, in certain situations. Therefore, instable features are more likely to indicate stylistic choices of the author (548).

Como se infiere de esta explicación, el análisis de la inestabilidad sería mucho más aparatoso que el de la frecuencia y no corresponde llevarlo a cabo aquí. Por eso en este trabajo el único factor determinante será la frecuencia. Seleccionar las palabras más frecuentes de un texto implicará inevitablemente recoger una ingente cantidad de palabras estructurales. No obstante, teniendo en cuenta que la noción de palabra que se está manejando es, dentro de la clasificación de Bloomfield (1933), la ortográfica, cierto

es que la extracción de estas palabras no diferencia entre distintos significados que corresponden a un mismo significante. Por eso, cuando lo interesante radica en recoger al mismo tiempo información contextual se recurre a los n-gramas de palabras, conjuntos de un determinado número (n) de palabras, con la desventaja de que «the classification accuracy achieved by word n-grams is not always better than individual word features» (543). Del mismo modo existen los n-gramas de caracteres, cuya utilidad reside en la obtención de información morfológica de rango menor a la palabra y sintáctica de rango menor a la frase. En ambos procedimientos lo difícil de especificar es el valor adecuado de n, que depende de diversos factores: la lengua sobre la que se trabaje, el tipo de información que se desee recabar, los objetivos del experimento, etc. A partir de este trasvase cuantitativo del estilo de un texto puede comenzar la aplicación de métodos numéricos con el fin de iniciar una comparación de estilos que sirva para inferir conclusiones relativas a la autoría.

3. METODOLOGÍA

Se describen a continuación las estrategias seleccionadas para este trabajo y su funcionamiento. La primera, *Stylo*, sigue el enfoque por modelo, mientras que las siguientes se adscriben al enfoque por caso. Las estrategias empleadas disponen de una serie de parámetros que pueden adquirir diferentes valores. Configurarlas adecuadamente para el experimento es esencial para obtener resultados coherentes. No hay ninguna instrucción que explicite la configuración que mejor va a funcionar en cada experimento, por lo que es necesario realizar antes unas cuantas pruebas con un corpus de autorías conocidas y fijar, para cada estrategia, los valores que arrojen los resultados más apropiados. En este caso el corpus de entrenamiento (*training corpus*) está constituido por una serie de obras de autoría conocida que serán después integradas en el corpus de prueba (*test corpus*), cuya constitución se detalla en el apartado 4.2.

3.1 *Stylo* de R

Stylo es un paquete del software libre R, el entorno de programación en código abierto de referencia para el análisis estadístico. Sus creadores lo definen como un «flexible R package for the high-level stylistic analysis of text collections» (Eder, Rybicki y Kestemont, 2016: 107). *Stylo* fue desarrollado específicamente para desempeñar tareas relacionadas con el análisis estilométrico de textos, por lo que ofrece una intuitiva GUI (*graphical user interface*) para usuarios que no estén familiarizados con los lenguajes de programación, que se activa con el comando `stylo()`. El proceso que sigue es, básicamente: «(i) textual data is acquired, (ii) the texts are preprocessed, (iii) stylistic features are extracted, (iv) a statistical analysis is performed, and finally, (v) the results are outputted (e.g. visualized)» (109). El preprocesamiento de los textos se refiere, en esencia, a su tokenización, es decir, «the process of dividing a string of input texts into countable units, such as word tokens» (110). El concepto token «representa la idea intuitiva de que una palabra es una cadena de caracteres alfabéticos rodeados de espacio» (Calvo Tello, 2016: 150). Los «stylistic features» que *Stylo* es capaz de extraer son n-gramas de palabras y de caracteres, rasgos ambos que «have been listed among the most effective stylistic features in survey studies in the field» (Eder, Rybicki y Kestemont, 2016: 111). Para diseñar un análisis estadístico de esas «features» habrá que centrarse en los ítems de alta frecuencia. Por defecto, *Stylo* emplea 1-gramas de palabras

y calcula las 100 MFW (*more frequent words*). La elección de ese valor se debe a que «in the earlier studies, sets of at most 100 frequent words were considered adequate to represent the style of an author» (Stamatos, 2009: 542). Los tres principales análisis que ejecuta son el *Principal Component Analysis* (PCA), bidimensional, el *Bootstrap Consensus Tree* (BCT), de apariencia radial, y el *Cluster Analysis* (CA), que será el empleado aquí y que consiste en la agrupación de los textos según similitudes estilísticas, lo que equivale, si los ajustes son correctos, a su agrupación por autor. Las opciones de visualización de R en cuanto a títulos, colores, forma de los gráficos, formato de archivo, etc. son amplias y permiten obtener resultados fácilmente interpretables sin un conocimiento profundo del software.

Para llevar a cabo los análisis estadísticos que *Stylo* posibilita se emplea la medida Delta, creada por John Burrows en 2002 específicamente con fines estilométricos. Su funcionamiento consiste en:

> First, this method calculates the *z*-distributions of a set of function words (originally, the 150 most frequent words). Then, for each document, the deviation of each word frequency from the norm is calculated in terms of *z*-score, roughly indicating whether it is used more (positive *z*-score) or less (negative *z*-score) times than the average. Finally, the Delta measure indicating the difference between a set of (training) texts written by the same author and an unknown text is the mean of the absolute differences between the *z*-scores for the entire function word set in the training texts and the corresponding *z*-scores of the unknown text (Stamatos, 2009: 554).

Por consiguiente, cuanto más bajo el valor de Delta más alta la similitud entre los textos estudiados. Recientemente, Maciej Eder (2016) ha realizado algunas innovaciones sobre el método clásico de Burrows. En su experimento, Calvo Tello (2016: 169) comprueba que con estas reformas se obtienen mejores resultados. *Stylo* permite escoger tanto la medida de Burrows como la de Eder e incluso otras que se basan en métodos matemáticos diferentes.

Stylo ofrece dos vías posibles para llevar a cabo la experimentación: la que facilita la GUI y la manual. En la reproducción del experimento con las obras de autoría desconocida se profundizará en la segunda explicitando los valores que cada argumento debe tomar y se omitirá la primera por su brevedad y sencillez. El primer paso, en ambas, consiste en cargar el paquete *stylo* con el comando `library(stylo)` y fijar como directorio de trabajo la carpeta donde se almacenen los textos con el comando `setwd(/nombredeldirectorio)`.

3.1.1 GUI (*graphical user interface*)

Basta introducir el comando `stylo()` para que aparezca en pantalla la GUI, que se desglosa en las pestañas *input & language, features, statistics, sampling y output.* Por defecto aparece seleccionada una serie de opciones que habrá que alterar ligeramente. Los cambios que hay que realizar son:

→ En la pestaña *input & language*, marcar la codificación UTF-8, que es la que sigue el corpus del trabajo y establecer como lengua el español.

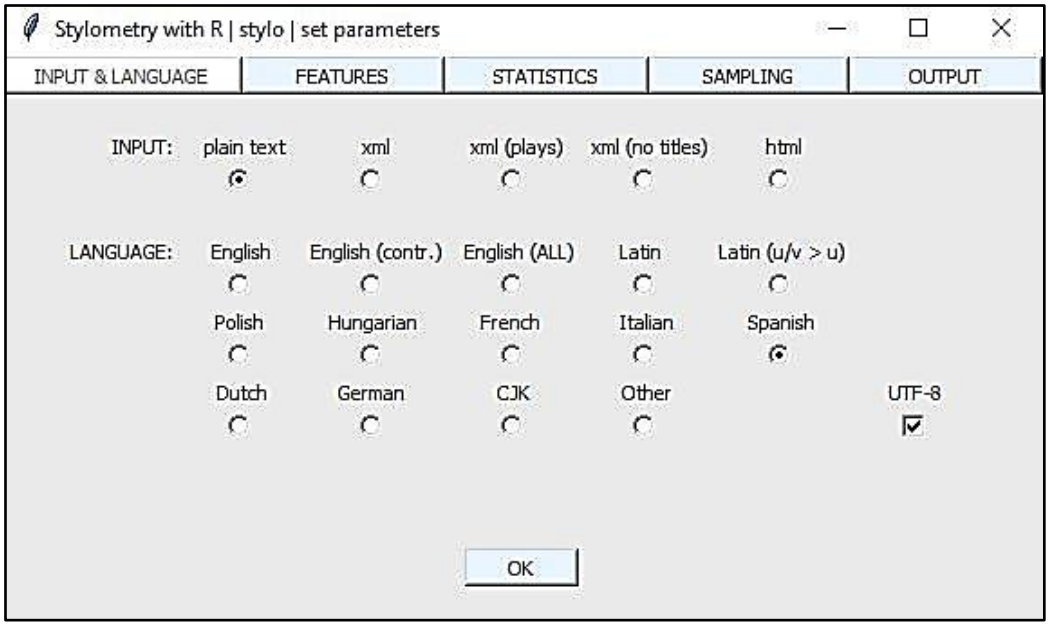


Figura 1. Pestaña *Input & Language* con la configuración pertinente

→ En la pestaña *features* desmarcar la casilla *preserve case* para poner en minúscula todos los caracteres del texto (de otro modo un mismo token en mayúscula o minúscula se interpretará como dos diferentes) y seleccionar el valor de MFW (*most frequent words*) deseado. De esto dependerán básicamente los resultados. Se mantiene la selección de 1-gramas de palabras y no se le da ningún valor a *culling*.

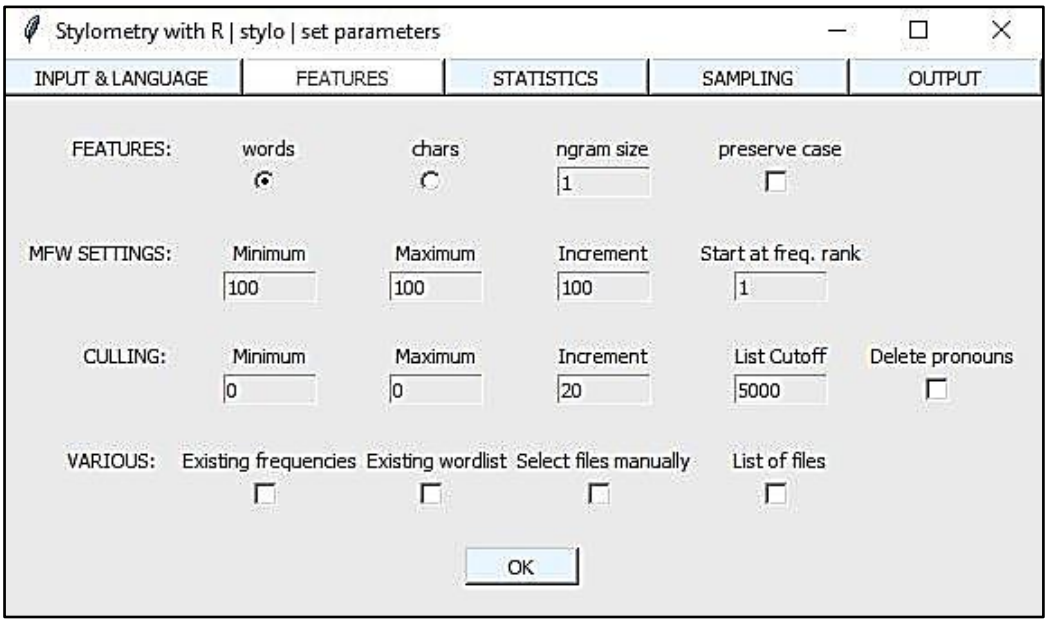


Figura 2. Pestaña *features* con la configuración pertinente

→ En la pestaña *output*, señalar los ajustes gráficos deseados: visualización del dendograma en diversos formatos de archivo, tamaño del gráfico, colores, títulos, etc.

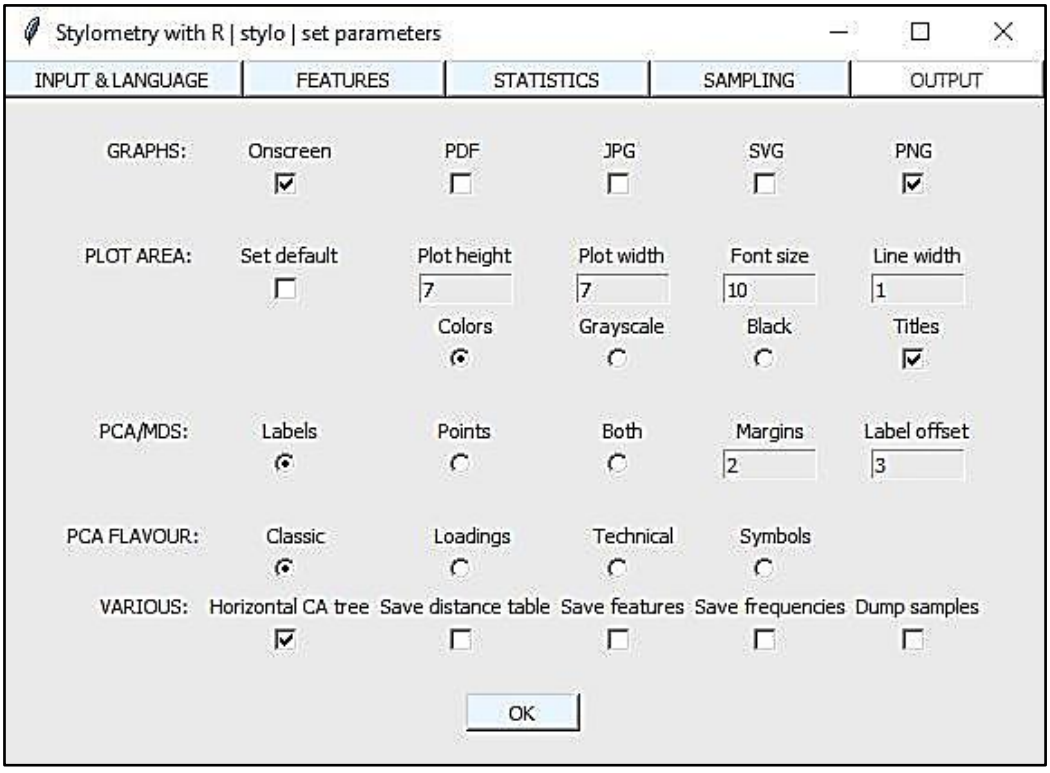


Figura 3. Pestaña *output* con la configuración pertinente

Una vez realizados estos cambios, se puede ejecutar el análisis. R devolverá un dendograma con los resultados que se interpreta de manera muy intuitiva.

3.1.2 Método manual

La forma más sencilla de ejecutar correctamente los comandos es seguir los pasos que ofrecen como ejemplo Maciej Eder, Jan Rybicki y Mike Kestemont en su artículo «Stylometry with R: A Package for Computational Text Analysis» (2016), que se resumen en la siguiente tabla:

Acción	Comando
Cargar el corpus	**`load.corpus`** `(files, corpus.dir, encoding)`
Tokenizar el corpus, es decir, dividir cada texto en palabras	**`txt.to.words.ext`** `(language, preserve.case)`
Extraer n-gramas del corpus	**`txt.to.features`** `(ngram.size, features)`
Seleccionar los x n-gramas más frecuentes	**`make.frequency.list`** `(head)`
Combinar las frecuencias relativas de los n-gramas extraídos y seleccionados en una matriz	**`make.table.of.frequencies`** `(features)`
Realizar el análisis estadístico y visualizar los resultados	**`stylo`** `(frequencies, analysis.type, mfw.max, mfw.min, gui, write.png.file, custom.graph.title)`

Tabla 1. Acciones que lleva a cabo R con los respectivos comandos que las ejecutan

Cada comando requiere una serie de argumentos, recogidos entre paréntesis, que permiten ajustar los parámetros de frecuencias, n-gramas, lengua, tipo de análisis, codificación, etc. deseados. En todo caso, lo más relevante es, al igual que en la GUI, el valor de MFW que se le ordene seleccionar al programa. Según la cita antes reproducida de Stamatos (2009: 542), a MFW debería ajustársele como máximo el valor 100. Sin embargo, siguiendo otros experimentos consultados en la comentada bibliografía, un valor tan pequeño podría no recoger resultados relevantes porque las palabras más frecuentes serán las mismas en todos los textos (las más frecuentes de la lengua: y, que,

de, en, a, por, etc.). También es cierto que un valor demasiado alto recogerá palabras cuya presencia en el texto no sea tan frecuente como para tener poder discriminatorio, por lo que habrá que buscar un valor intermedio. Se muestran aquí los dendogramas obtenidos para resultados de 100, 250 y 500 MFW:

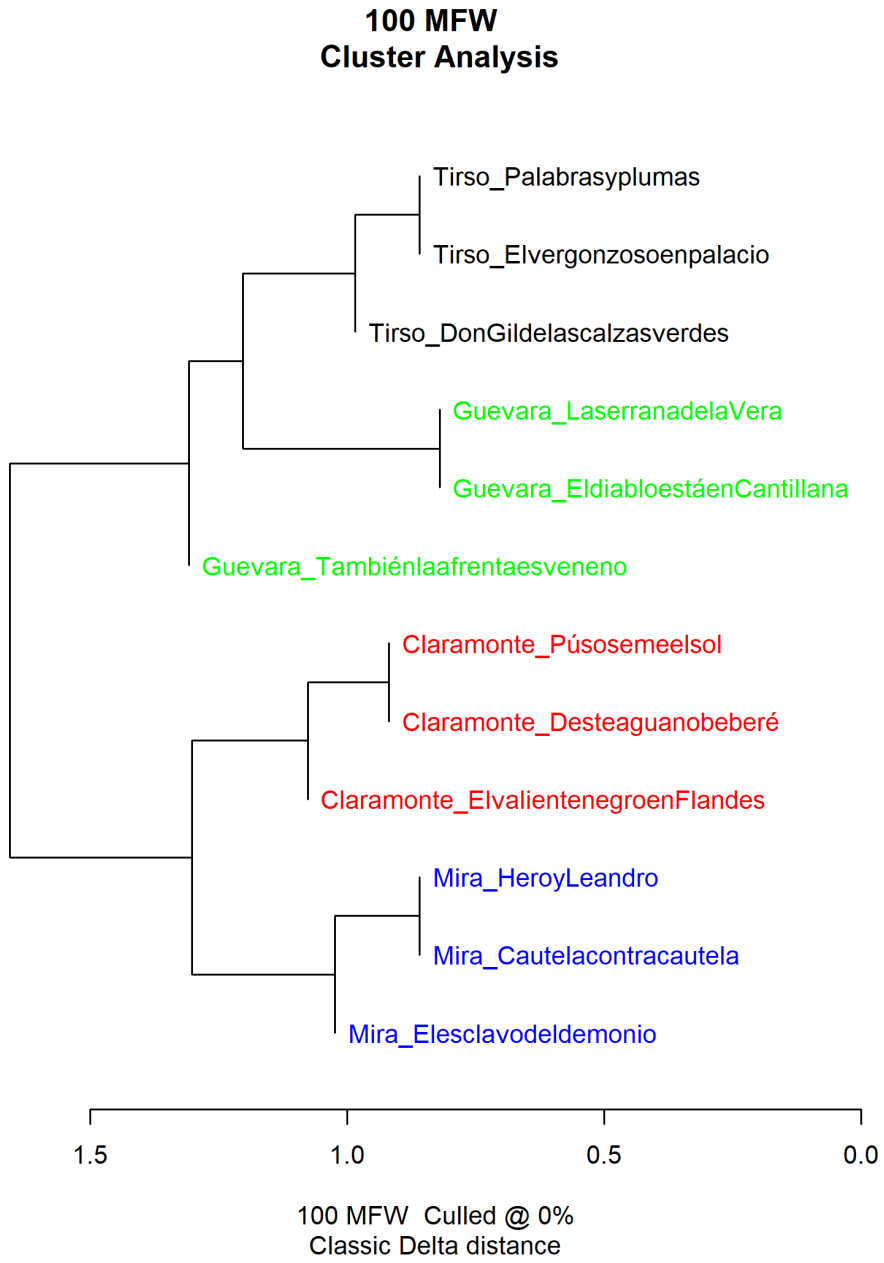


Figura 4. Dendograma del corpus de entrenamiento con los 100 tokens más frecuentes

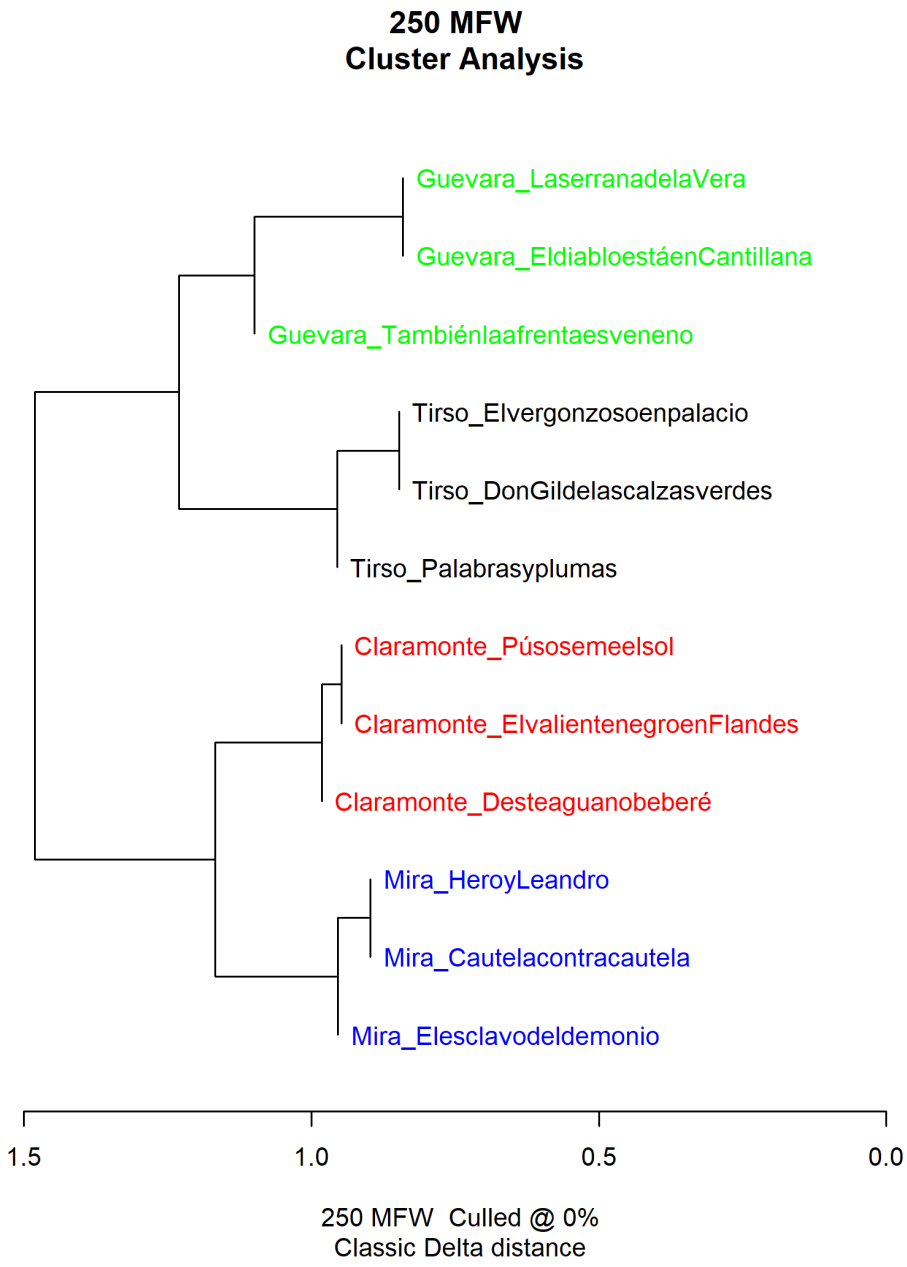


Figura 5. Dendograma del corpus de entrenamiento con los 250 tokens más frecuentes

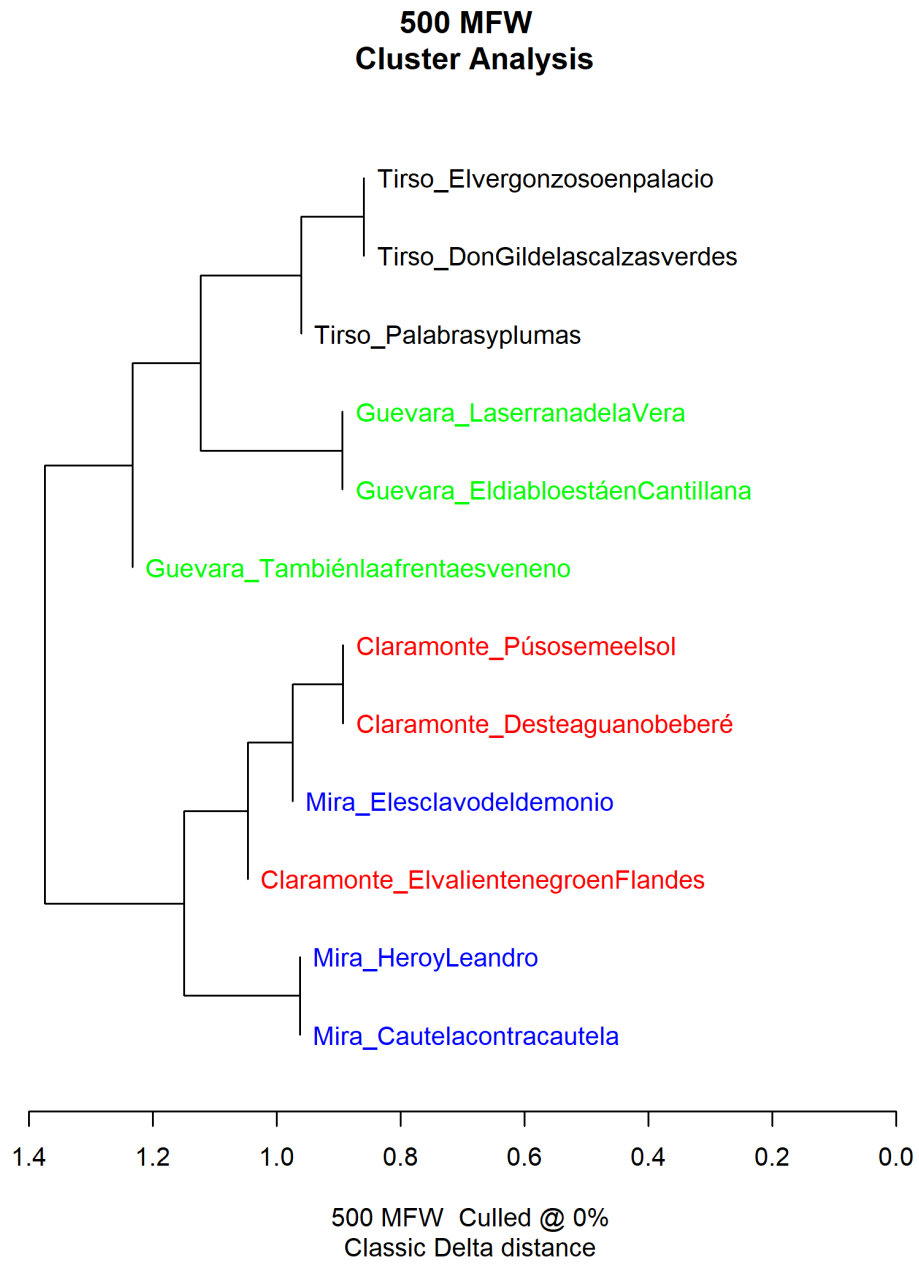


Figura 6. Dendograma del corpus de entrenamiento con los 500 tokens más frecuentes

Las tres configuraciones arrojan resultados más o menos adecuados: la de 100 MFW no agrupa correctamente la obra de Guevara *También la afrenta es veneno*, que según el dendograma podría adjudicarse también a Tirso. La de 500 MFW hace lo mismo y además extravía una obra de Mira agrupándola entre las de Claramonte. Por último, la de 250 MFW es la más certera, ya que es la única que no yerra en ninguna obra, por lo que será esta la empleada para el experimento posterior.

3.2 Medidas de distancia

Estas medidas se basan en distintos métodos matemáticos que realizan una serie de cálculos para llegar a un resultado numérico final que es interpretable al compararlo con los resultados obtenidos por las otras medidas. No fueron diseñadas, a diferencia de *Stylo*, específicamente con el fin de servir a los propósitos de los NTAAS, sino que se emplean habitualmente en otras tareas computacionales. Para testarlas se empleará la comedia de Tirso *Los lagos de San Vicente*, cuya distancia se medirá con respecto a todos los autores del corpus para comprobar que, efectivamente, pertenece a Tirso.

3.2.1 Divergencia Kullback-Leibler

La divergencia Kullback-Leibler, desarrollada por Solomon Kullback y Richard Leibler en 1951, es una medida de distancia que se basa en una ecuación que compara las distribuciones de frecuencia de los elementos de dos (para más de dos será necesario hallar la media de los sucesivos resultados) corpus diferentes. Álvaro Iriarte, Pablo Gamallo y Alberto Simões en su experimento para desarrollar estrategias automáticas para buscar especificidades lexicales dentro de conjuntos de textos mediante lemas y MWE (*multiple word extraction*) concluyen que es «uma medida robusta» (2018: 24) porque extrae los valores esperados. De sus dos hipótesis de partida —que la divergencia Kullback-Leibler servirá para comparar textos no anotados previamente y que la MWE arrojará mejores resultados que el uso de palabras simples— solo la primera se verifica con éxito.

3.2.2 *Perplexity* y *ranking*

Perplexity y *ranking* son dos medidas que tienen en común que pueden estar basadas tanto en n-gramas de palabras como en «n-gram models of characters extracted from text corpora» (Gamallo, Pichel y Alegría, 2017: 153) y su empleo en tareas de identificación y distancia entre lenguas. La ventaja de emplear n-gramas de caracteres radica en que contienen un alto grado de información fonológica, morfológica y sintáctica, frente a otros métodos que se basan exclusivamente en el léxico. He aquí un ejemplo de cómo un n-grama de caracteres con un valor suficientemente alto de n encierra información sintagmática:

> The 7-gram *ion#de#* (where‘#’ represents a blank space) is a frequent sequence of letters shared by several Romance languages (e.g. French, Spanish, or Galician). This 7-gram might be considered as an instance of the generic pattern ‘‘*noun-prep-noun*’’ since *ion* is a noun suffix and *de* a very frequent preposition (153).

Perplexity es una medida de «how well a model fit the test data» (155). Se define como «the inverse probability of the test text given the model» (153). También se usa para medir la calidad de modelos del lenguaje dado un conjunto de textos de prueba. En este trabajo se extrapolará su utilidad al ámbito de las atribuciones de autoría, siendo los modelos autores y los textos de prueba comedias. Empleará, como en el experimento de Pablo Gamallo, José Ramom Pichel e Iñaki Alegría, 7-gramas de caracteres. Al ser una medida de distancia, cuanto más bajo sea su valor, más alto será el parecido entre dos textos.

Ranking es una medida de frecuencia cuyo funcionamiento se resume en:

> N-grams are ranked according to frequency in a training corpus, and those with highest frequencies are selected while the rest are discarded. This gives us the pruned character n-grams profile for each language. A *language profile* is thus the ranked list of the most frequent n-grams in the training corpus. Unlike n-gram language models, language profiles do not make use of prior probabilities but simply of ranked lists. The ranking-based distance between two languages is obtained by comparing the ranked lists of the two languages (155).

Para comprender su aplicación en este trabajo basta sustituir *language* en la cita anterior por autor y matizar que en este experimento emplea n-gramas de palabras (concretamente unigramas) y no de caracteres, puesto que en el experimento de Pablo Gamallo, Marcos García, Susana Sotelo y José Ramom Pichel (2014) sobre la detección del idioma en tweets «los modelos con unigramas de palabras funcionan mejor que el uso de n-gramas de caracteres» (12).

En el experimento de Gamallo, Pichel y Alegría (2017) son evaluadas ambas medidas y *perplexity* es la que obtiene los resultados más adecuados.

3.2.3 Similitud coseno

El coseno es una medida de similitud integrada en los algoritmos de multitud de buscadores de información en línea. También es empleada en otros sistemas de recuperación de la información y minería de datos. El primer paso para obtenerla es «assigning a weight for each term in a document» (Manning, 2008: 133), de manera que «a document may be viewed as a vector of term weights». El peso que se le asigna a cada término viene dado por la frecuencia relativa de ese término a lo largo del documento. Así se consigue convertir textos en vectores. Normalmente esto se aplica a

una *query* (la consulta que introducimos en un buscador) y un documento, pero en este trabajo se aplicará entre dos documentos, ya que el método es exactamente el mismo:

> By viewing a query as a "bag of words", we are able to treat it as a very short document. As a consequence, we can use the cosine similarity between the query vector and a document vector as a measure of the score of the document for that query (124).

Una vez se han obtenido los vectores correspondientes a los textos, «a plausible scoring mechanism then is to compute a score that is the sum, over the query terms, of the match scores between each query term and the document» (117) o, en este caso, entre cada pareja de textos. Entonces la similitud coseno se calcula dividiendo el producto escalar de los vectores de los textos entre el producto de los módulos de esos vectores (121). Gráficamente, la medida es el valor del coseno del ángulo que forman los vectores. Al ser una medida de similitud, cuanto más alto sea su valor, más alto será el parecido entre los textos, siendo el máximo valor 1 (valor máximo del coseno de dos ángulos cualesquiera). Funciona, por tanto, de manera inversa a las otras tres, que son medidas de distancia.

3.2.4 Ejecución de las cuatro medidas

Las cuatro medidas empleadas (tres de distancia y una de similitud, coseno) no necesitan una configuración previa para poder recibir datos. Se encuentran compiladas en un ejecutable escrito en lenguaje PERL, específicamente diseñado para este trabajo, que se debe lanzar desde la terminal de comandos de cualquier distribución del sistema operativo Linux, al que se puede acceder desde un computador Windows creando una máquina virtual con el software de Oracle VM VirtualBox. Una vez dentro hay que conectarse al servidor que tiene almacenados los datos, situarse en el directorio en el que están guardados los textos y ejecutar el comando `sh run.sh Nombredeltextoquesequieracomparar.txt`. El cálculo de todas las medidas suele llevar medio minuto y los resultados finales se almacenan en diferentes archivos clasificados por medida.

Para probar la funcionalidad de esta estrategia se ha contrastado la comedia de Tirso *Los lagos de san Vicente*, y los resultados han sido los siguientes:

Perplexity	Kullback-Leibler	*Ranking*	Coseno
3,2725 Tirso	1,0585 Guevara	0,499 Tirso	0,8587 Tirso
3,364 Mira	1,0589 Mira	0,571 Mira	0,8528 Guevara
3,5119 Claramonte	1,0593 Claramonte	0,577 Claramonte	0,8508 Mira
3,7972 Guevara	1,0815 Tirso	0,601 Guevara	0,8505 Claramonte

Tabla 2. Resultados obtenidos con las cuatro medidas de distancia/similaridad al confrontar la obra de Tirso *Los lagos de san Vicente* con el corpus de entrenamiento

El primer reajuste que cumple realizar es la normalización de los valores según la fórmula:

$$\frac{x - Min}{Max - Min}$$

$x =$ valor original para un autor
$Min =$ valor más bajo de los obtenidos para los cuatro autores
$Max =$ valor más alto de los obtenidos para los cuatro autores

Esta normalización es necesaria para que posteriormente todas las medidas tengan el mismo peso a la hora de calcular su media, ya que, como se ve en la tabla 2, los valores absolutos originales se mueven en rangos distintos, lo que impide su comparación. Se aprecia en todas las medidas que la diferencia entre el valor máximo y el mínimo es muy pequeña, lo que corrobora la naturaleza similar de los textos tratados. Como se deduce de la fórmula, al autor con el estilo más próximo al del texto que se está cotejando se le asignará el valor redondo 0,0 y al que posea el estilo más alejado, el valor 1,0. Los resultados normalizados son:

Perplexity	Kullback-Leibler	*Ranking*	Coseno
0,0000 Tirso	0,0000 Guevara	0,0000 Tirso	1,0000 Tirso
0,1744 Mira	0,0174 Mira	0,7059 Mira	0,2805 Guevara
0,4563 Claramonte	0,0348 Claramonte	0,7647 Claramonte	0,0366 Mira
1,0000 Guevara	1,0000 Tirso	1,0000 Guevara	0,0000 Claramonte

Tabla 3. Resultados normalizados obtenidos con las cuatro medidas de distancia/similaridad al confrontar la obra de Tirso *Los lagos de san Vicente* con el corpus de entrenamiento

Con el fin de que la visualización conjunta de los resultados sea más clara es preciso realizar un reajuste adicional: calcular el inverso de los valores de la medida coseno para interpretarla como una medida de distancia más. Para ello se le resta a 1 el valor obtenido para cada autor.

Coseno
0,0000 Tirso
0,7195 Guevara
0,9634 Mira
1,0000 Claramonte

Tabla 4. Reajuste de los resultados normalizados obtenidos con la similitud coseno al confrontar la obra de Tirso *Los lagos de san Vicente* con el corpus de entrenamiento

Comparando los valores por medida se aprecia que el consenso entre ellas es considerable salvo por Kullback-Leibler, que es la que más se aleja de las demás al colocar al autor verdadero de la comedia en último lugar. Para no generar un exceso de datos numéricos que dificulten la interpretación de los resultados y teniendo en cuenta que, en general, estas medidas no son tan robustas como la distancia Delta que emplea *Stylo*, se combinan los resultados de las cuatro haciendo una media aritmética. El valor obtenido, entre 0 y 1, es el que determina lo parecidos que son los textos y por tanto sirve para realizar comparaciones entre los autores. Cuanto más próximo a 0, menos distancia, es decir, más parecido. Cuanto más próximo a 1, más distancia, es decir, menos parecido.

Tirso_LoslagosdeSanVicente
0,250 Tirso
0,465 Mira
0,564 Claramonte
0,680 Guevara

Tabla 5. Media aritmética de los resultados normalizados obtenidos con las cuatro medidas al confrontar la obra de Tirso *Los lagos de san Vicente* con el corpus de entrenamiento

Como era previsible, la adjudicación a Tirso de Molina realizando la media de las cuatro distancias es rotunda, lo que demuestra el potencial de estas herramientas y la utilidad de su fusión. En el experimento posterior los resultados también servirán para determinar cuál es la mejor medida para casos de atribuciones de autoría como este. Todavía hay una tercera manera de exhibir los resultados, más adecuada y fácil de interpretar. Se trata de calcular la media de las posiciones que los autores obtuvieron en los resultados de cada una de las medidas. Al ser cuatro las medidas, el valor de esta nueva estrategia estará comprendido entre 1 (el autor quedó en primera posición en los resultados de las cuatro medidas) y 4 (el autor quedó en última posición en los resultados de las cuatro medidas). Se muestran las posiciones obtenidas en la siguiente tabla y su media calculada en la última columna:

	Perplexity	**Kullback**	***Ranking***	**Coseno**	**Media de las posiciones**
Tirso	1	4	1	1	**1,75**
Mira	2	2	2	3	**2,25**
Guevara	4	1	4	2	**2,75**
Claramonte	3	3	3	4	**3,25**

Tabla 6. Posiciones de los autores según los resultados obtenidos con cada medida al confrontar la obra de Tirso *Los lagos de san Vicente* con el corpus de entrenamiento y media aritmética de estas

En el experimento posterior se ofrecerá el valor de coseno directamente reajustado como distancia. Del mismo modo, los resultados se presentarán ya normalizados. Se desglosarán en estas tres formas de presentación (por medida como en la tabla 3, por media de las medidas como en la tabla 5 y por media de las posiciones como en la tabla 6), concediéndole primacía a la última por su transparencia.

4. APLICACIÓN DE LAS ESTRATEGIAS AL TEATRO DE TIRSO DE MOLINA

En este apartado se desarrollan los experimentos diseñados con las estrategias antes explicadas y se analizan los resultados obtenidos. Para ello se ofrece previamente un estado de la cuestión del debate tirsiano en 4.1, que pretende ser un compendio de la extensa bibliografía disponible sobre el tema, y se detalla el embrollado proceso de configuración del corpus en 4.2. Después de especificar los pormenores de los experimentos con cada una de las estrategias y presentar los resultados en el subapartado 4.3, se realiza una discusión conjunta de los mismos en 4.4 que permite extraer conclusiones relevantes.

4.1 Problemas de autoría en el teatro de Tirso de Molina. Estado de la cuestión

La elección de la producción de Tirso de Molina (1579-1648), seudónimo del fraile mercedario Gabriel Téllez y uno de los dramaturgos más prestigiosos del siglo XVII, se debe a que los problemas de atribución de autoría que rodean a varias de sus obras más célebres son probablemente el enigma autoral más representativo de toda la historia de la literatura española después de los dos interrogantes críticos por excelencia: el autor del *Lazarillo* y la identidad de Avellaneda. Ambas incógnitas han sido recientemente examinadas a la luz de técnicas de atribución de autoría no tradicionales con resultados más o menos satisfactorios, pero desde luego avanzados con respecto a los inferidos del empleo de métodos convencionales. Javier de la Rosa y Juan Luis Suárez (2016) obtienen resultados que apuntan al jurista Juan Arce de Otálora para el *Lazarillo*, mientras que Javier Blasco (2016) asocia el *Quijote* apócrifo a figuras como Castillo Solórzano o el propio Tirso. Es cuando menos sorprendente que el asunto del teatro de Tirso todavía no haya sido estudiado con estas herramientas si se tiene en cuenta la dilatada bibliografía que existe sobre el tema y la perseverancia de los investigadores implicados en el debate, que han reafirmado a lo largo de los años insistentemente sus posturas. Las repercusiones que conllevaría la resolución definitiva del enigma serían muy significativas, puesto que acarrearían la desestimación de la posición de un autor fundamental en el canon, ya que las comedias cuestionadas son las que más atención han recibido por parte de la crítica y más veces han sido llevadas a escena.

La edición y difusión de la obra tirsiana es un ejemplo transparente del intrincado proceso de transmisión textual de las comedias del siglo de Oro, que se

imprimían bien en partes (generalmente de doce comedias cada una), bien en sueltas, tras haber sido representadas y pasando por las manos de, por lo menos, el director y el editor, sin que el poeta tuviera control alguno en ninguna etapa del procedimiento. No es de extrañar, por lo tanto, que durante el proceso surgieran equívocos. Además, en el caso de dramaturgos reputados hay que sumar otro problema ecdótico, y es que «la notoriedad de su obra hace que varios avispados editores publiquen fraudulentamente comedias ajenas como suyas» (Rodríguez López-Vázquez, 1990: 6), fenómeno que también era habitual con Lope y Calderón, por ejemplo.

Blanca Oteiza identifica las cuatro fuentes a través de las cuales nos ha llegado la obra del mercedario: «manuscritos, pocos; de las ediciones de las cinco *Partes* las más; en sueltas; y en sueltas integradas en volúmenes colectivos» (2000: 100). En su artículo enumera varios de los errores de los que adolecen las ediciones: omisión de versos, cambios en el orden de los versos, saltos de palabras dentro de un verso rompiendo la medida, errores de puntuación y malas lecturas. De las cinco partes que fueron publicadas bajo el nombre de Tirso, la crítica advierte de la no fiabilidad de dos de ellas, la segunda y la tercera, editadas en la década de 1630. En el caso de la segunda, es el propio Tirso el que pone sobre aviso de la fraudulenta maniobra editorial: «Dedico, destas doce comedias, cuatro que son mías en mi nombre, y en el de los dueños de las otras ocho (que no sé por qué infortunio suyo, siendo hijas de tan ilustres padres las echaron a mis puertas), las que restan» (Oteiza, 2000: 106). Desafortunadamente, Tirso no informa de cuáles son las ocho obras que no salieron de su pluma. A día de hoy, la crítica ha sido capaz de dilucidar la autoría de varias, siendo la situación en torno a la *Parte Segunda* de Tirso la siguiente:

Comedia	**Autor**
La reina de los reyes	Hipólito de Vergara
Amor y celos hacen discretos	Tirso de Molina
Quién habló pagó	Rodrigo de Herrera
Siempre ayuda la verdad	Antes: ¿Luis Belmonte Bermúdez? ¿Juan Ruiz de Alarcón? ¿Rodrigo de Herrera? ¿Tirso de Molina?
	Desde abril por García-Reidy (2019): Lope de Vega

Los amantes de Teruel	Juan Pérez de Montalbán
Por el sótano y el torno	Tirso de Molina
Cautela contra cautela	Mira de Amescua
La mujer por fuerza	¿Tirso de Molina?
El condenado por desconfiado	¿Mira de Amescua? ¿Andrés de Claramonte? ¿Vélez de Guevara? ¿Tirso de Molina? ¿Colaboración?
Primera parte de don Álvaro de Luna	Mira de Amescua
Segunda parte de don Álvaro de Luna	Mira de Amescua
Esto sí que es negociar	Tirso de Molina

Tabla 7. Comedias de la *Parte Segunda* de Tirso con sus respectivos autores o candidatos más probables (entre signos de interrogación)

En resumen, la principal duda se asienta sobre la cuarta comedia que sí pertenece a Tirso, siendo la candidata más probable *La mujer por fuerza*, pero sin poder descartar *El condenado por desconfiado* (y hasta hace muy poco, *Siempre ayuda la verdad*). Para esta se han propuesto multitud de argumentos que desestiman la autoría de Tirso y se la otorgan a otros autores, pero el desacuerdo entre la crítica es extremo e incluso se baraja la hipótesis de la colaboración: «Claramonte sería el autor del primer acto, Mira el del segundo, y el tercero sería obra conjunta, o tal vez de otro autor, como Vélez o Belmonte, colaborador habitual de Mira de Amescua» (Rodríguez López-Vázquez, 2010: 144). La colaboración también ha sido propuesta para *La venganza de Tamar*, presente en la *Parte Tercera* y en la que probablemente interviniese la mano de Calderón. Sin embargo, los problemas de autoría más importantes conciernen a las ediciones sueltas.

Entre las que más polémica han suscitado se encuentra *El rey don Pedro* o *El infanzón de Illescas*, atribuida a Tirso desde el siglo XIX por una «conjetura crítica desprovista de base documental» (Rodríguez López-Vázquez, 1990: 8) del poeta Juan Eugenio Hartzenbusch y hoy oscilando entre Lope, Calderón y Andrés de Claramonte. Más aclarada parece estar la autoría de *La ninfa del cielo*, que ha sido restituida a Luis Vélez de Guevara en la edición crítica de Alfredo Rodríguez López-Vázquez (2008) para Cátedra sin ninguna refutación hasta el momento. Pero la obra que más

controversia ha generado por su excepcional influencia ha sido la archiconocida *El burlador de Sevilla*.

> *El burlador de Sevilla*, cuya edición príncipe, atribuida a Tirso, apareció integrada en *Doce comedias nuevas de Lope de Vega Carpio y otros autores. Segunda parte* (Barcelona, Gerónimo Margarit, 1630), y cuya autoría sigue siendo cuestión debatida hoy: desde la atribución a Claramonte por Alfredo Rodríguez a Luis Vázquez, que la atribuye a Tirso sin dudar, hay posturas de todos los matices; o de *Tan largo me lo fiáis*, relacionada estrechamente con *El burlador*, que salió suelta a nombre de Calderón, sin datos de imprenta, aunque parece ser de Sevilla, Francisco Lira, hacia 1635 (Oteiza, 2000: 112).

El núcleo del problema, pues, reside en la relación genética entre el texto del *Burlador* que ha venido siendo editado tradicionalmente y la versión alternativa de *Tan largo me lo fiáis*, coincidentes en «1433 versos, que representan más de un 60% del total» (García Gómez, 2005: 282). En este punto la crítica se escinde en dos posturas opuestas:

- Xavier A. Fernández, apoyado por P. Guenoun (1962), J. Casalduero (1977) y J. M. Ruano de la Haza (1995), ha defendido la primacía de composición de *El burlador* y la existencia de un texto primitivo perdido que «se transmite independientemente tanto a *El burlador* como a *Tan largo*» (García Gómez, 2005: 282).
- Alfredo R. López-Vázquez, apoyado por R. J. Mayberry (1962), G. E. Wade (1962), A. E. Sloman (1965), M. R. Lida de Malkiel (1966) y D. Rogers (1977), ha defendido la primacía de composición del *Tan largo*, que habría sido refundido posteriormente en *El burlador*.

José María Ruano de la Haza, a pesar de situarse en la hipótesis de primacía de *El burlador* aporta una perspectiva mucho más singular poniendo el foco sobre el personaje del gracioso, Catalinón, cuyas apariciones coinciden con los fragmentos que tienen en común ambos textos:

> El autor del texto original vende su manuscrito a una compañía; esta compañía lo representa en exclusiva absoluta durante varios años; un actor de esta compañía, probable pero no necesariamente el que durante esos años hizo el papel de Catalinón, se pasa a otra compañía, llevándose consigo el traslado parcial de su «papel» que incluía los «pies de entrada» y probablemente el «papel» de Aminta, y, en su memoria, varios pasajes importantes de la comedia junto con su traza e historia. Con ayuda del poeta, o autor, de la otra compañía, y basándose en el manuscrito parcial que lleva consigo, reconstruye una versión de la comedia (1995: s. p.).

Para Ruano de la Haza, esa versión reconstruida sería el *Tan largo* y ese poeta Andrés de Claramonte. Sin embargo, esta teoría da una vuelta de tuerca cuando en 2005 Ángel García Gómez encuentra en un cartapacio cordobés la prueba de una

representación de una comedia titulada *Tan largo me lo fiáis* en 1617 (la primera representación de *El burlador* de la que se tiene constancia fue en 1625), lo que refuerza la hipótesis de primacía del *Tan largo*.

Rodríguez López-Vázquez atribuye la autoría del *Tan largo* (y en consecuencia la de *El burlador*) al dramaturgo murciano Andrés de Claramonte (c. 1560-1626), que se erige así en figura central en este debate, a pesar de que su nombre no resulte familiar en absoluto. La causa de la escasez de bibliografía sobre este autor, su ausencia de las historias de la literatura y la gran porción de su producción que permanece aún sin editar hay que rastrearla, según Rodríguez López-Vázquez, en dos hechos. El primero es su papel como «autor» teatral, es decir, director «que no publicó en vida comedias, tenía fama común de "plagiario", no es alabado por ningún poeta contemporáneo» (1990: 12). El segundo es la opinión que le mereció a Menéndez Pelayo, que fue esta: «dramaturgo vulgar y adocenado que se dedicó a la piratería literaria» (1983: 88). Este juicio personal condicionó negativamente las sucesivas aproximaciones a la figura de Claramonte. Luis Vázquez, por ejemplo, basa su argumentación en contra de Claramonte en que «un poeta mediocre no puede ser el autor del don Juan original» (1995: 184). De esta manera, la resolución del problema de autoría de *El Burlador* se vuelve esencial a la hora de determinar no solo el lugar que Claramonte debe ocupar en la historia de la literatura sino incluso su reputación. En palabras de Rodríguez López-Vázquez, «la consecuencia es la revalorización de la obra de Claramonte» (s. f.: 23). No obstante, Rodríguez López-Vázquez, que lleva casi toda su carrera intentado dilucidar este problema de autoría, emplea «argumentos demasiado complicados, confusos en ocasiones, y altamente subjetivos siempre» (Ruano de la Haza, 1995: s. p.) para atribuir a Claramonte no solo las dos versiones del primer don Juan, sino también *El rey don Pedro en Madrid*, *El condenado por desconfiado* y *La estrella de Sevilla*, esta última siguiendo a Sturgis E. Leavitt (1931) en su refutación de la atribución tradicional a Lope. Aunque él afirma que «no se trata de una investigación hecha para reivindicar el teatro de Claramonte» (s. f.: 24), este caso sirve como ejemplo de la actitud sesgada que prácticamente todos los académicos que se han enfrentado a este problema han manifestado, centrando su análisis en el elemento que más favorecía la hipótesis (o, en muchos casos, conjetura) que se proponían probar de entre todos los posibles: índices léxicos, metáforas recurrentes, métrica, nombres de personajes, contenido teológico, etc. y obteniendo siempre resultados insuficientes para demostrar rotundamente su teoría.

Por esta razón no han logrado llegar a un acuerdo en casi ningún punto y la solución al problema parece quedar todavía lejos.

4.2 Configuración del corpus

Teniendo en cuenta todo lo expuesto en el apartado anterior, lo ideal parece ser construir un corpus amplio con todas las obras de Tirso cuya autoría esté en disputa y varias obras representativas de todos los posibles autores. Sin embargo, los límites de tiempo, medios y extensión de este trabajo impiden semejante empresa. Para llevar a cabo cualquier experimento con técnicas de atribución de autoría no tradicionales es imprescindible disponer de los textos que van a ser analizados en texto plano (extensión .txt), por lo que deben estar adecuadamente digitalizados. El principal obstáculo que uno se encuentra al intentar reunir un corpus de comedias del siglo de Oro es que hay una carencia abrumadora de digitalizaciones incluso de los autores más reconocidos. La fracción de comedias auriseculares que se puede consultar de manera libre en línea es ínfima. La fuente en la que se alojan la mayoría es el portal de la Biblioteca Virtual Miguel de Cervantes (BVC), que ofrece multitud de textos literarios y críticos de toda la historia de la literatura española. Más focalizada hacia el género que nos ocupa está la página web de la Association for Hispanic Classical Theater (AHCT) con sede en Maryland, Estados Unidos, que presenta un corpus de comedias auriseculares de una variedad significativa de autores. Más allá de estas dos fuentes la tarea se vuelve prácticamente imposible salvo por algunos *e-books* preparados por particulares para editoriales concretas. Conviene destacar que el inconveniente de la falta de digitalizaciones es la consecuencia directa de otro problema mucho más grave: la escasez de ediciones. Gran cantidad de textos dramáticos del siglo de Oro permanecen a día de hoy sin editar, a pesar de que el acceso a la reproducción digital de los manuscritos ya es posible a través de portales como el Fondo Antiguo de la Universidad de Sevilla o la propia BVC. Este fenómeno es especialmente limitante en el caso del dramaturgo Andrés de Claramonte, al que la crítica ha prestado apenas atención, y dificulta enormemente la configuración de un corpus extenso para estudiar atribuciones. Por todo esto, el primer criterio que rige para la conformación de un corpus que sirva para examinar los problemas de autoría del teatro de Tirso de Molina será forzosamente la disponibilidad de los textos y el segundo las limitaciones que impongan las estrategias empleadas. En este caso las estrategias no infligen restricciones

considerables en cuanto a número de autores, obras, longitud de los textos, etc.; por lo tanto, será factible abarcar el siguiente conjunto de casos:

Texto	Posibles autores
El burlador de Sevilla	Claramonte / Tirso
El condenado por desconfiado	Amescua / Claramonte / Guevara / Mira / Tirso / Colaboración
La mujer por fuerza	Tirso / Otro (ningún nombre propuesto)
La ninfa del cielo	Guevara / Tirso
Tan largo me lo fiáis	Claramonte / Tirso

Tabla 8. Textos objeto de estudio del trabajo con sus respectivos posibles autores

Como se deduce de la tabla 8, los autores implicados en el estudio serán, a parte de Tirso, Andrés de Claramonte, Mira de Amescua y Luis Vélez de Guevara.

Las comedias del siglo de Oro son textos bastante parecidos entre sí, que responden a unos mismos esquemas dramáticos y estilísticos e incluso que han servido como inspiración unos de otros. En estas condiciones la discriminación entre autores se vuelve más complicada, por eso se deben recopilar textos que sean muy representativos del estilo de su autor. Para tomar una muestra representativa de las producciones de estos cinco autores más o menos equivalente en número de versos se ha primado de nuevo el criterio forzoso de la disponibilidad. En los casos en los que ese principio lo permitía (Tirso y Mira) se han seleccionado títulos cuya fecha de composición estuviese lo más próxima posible a la de las obras de autoría dudosa, respetando el principio de «all the texts per author should be written in the same period to avoid style changes over time» (Stamatos, 2009: 558). Se ha procurado la variedad temática, en especial en las elegidas de Mira. En el caso de Guevara no se han incluido las famosas *El diablo cojuelo* o *Los hermanos amantes* porque, al estar en prosa, podrían producir graves alteraciones en los resultados. Se ha empleado como fuente principal la BVC, de manera que casi todos los textos han sido editados en base a criterios similares. Cuando la comedia no estaba disponible en ese portal se ha recurrido a la AHCT o, en su defecto, a ediciones particulares. Finalmente, el corpus de prueba conformado es el que se detalla aquí (el corpus de entrenamiento más las obras de autoría en disputa):

Autor	Comedia	Fecha de composición	Fuente	Versos
Tirso de Molina	*El vergonzoso en palacio*	1610-1625	BVC	3968
	Don Gil de las calzas verdes	1615	BVC	3277
	Palabras y plumas	Anterior a 1627	BVC	3979
				11224
Mira de Amescua	*Cautela contra cautela*	Anterior a 1635	BVC	2845
	El esclavo del demonio	1612	BVC	3296
	Hero y Leandro	—	AHCT	3310
				9451
Andrés de Claramonte	*Deste agua no beberé*	Anterior a 1617	BVC	2743
	El valiente negro en Flandes	Anterior a 1638	Bubok	2973
	Púsoseme el sol	—	BVC	3206
				8922
Vélez de Guevara	*El diablo está en Cantillana*	1622	BVC	2621
	La serra de la Vera	1613	BVC	3306
	También la afrenta es veneno (acto primero)	—	BVC	1196
				7123
Desconocido	*El burlador de Sevilla*	1612-1617	BVC	2900
	El condenado por desconfiado	Anterior a 1635	BVC	2995
	La mujer por fuerza	Anterior a 1635	Clásicos Hispánicos	2886
	La ninfa del cielo	1610-1620	AHCT	3505
	Tan largo me lo fiáis	1612-1617	BVC	2760

Tabla 9. Obras que componen el corpus de prueba. Se intenta acotar la fecha de composición lo máximo posible

Una vez conseguidos los textos y pasados a texto plano se procede a su preprocesamiento. Para limpiar los textos de las impurezas que suelen contener se ha empleado una serie de comandos en la terminal de Ubuntu. Concretamente, las modificaciones operadas sobre los textos han sido su codificación en UTF-8, la sustitución de los saltos de línea de Windows por los de Linux (que no implican retorno de carro) y la supresión de la numeración de los versos. No se han excluido los nombres de los personajes antes de cada intervención porque la onomástica ha sido uno de los argumentos que los investigadores han aducido en sus distintas propuestas de atribuciones. El último paso antes de comenzar a usar las herramientas es la correcta catalogación de los textos, que deberán ubicarse en el mismo directorio y cuya nomenclatura deberá seguir el patrón «Autor_Titulo.txt» dado que *Stylo* emplea esa pauta para clasificar las obras de un mismo autor en base a un código cromático. Solo después de todo este proceso es viable iniciar la experimentación con los textos.

4.3 Diseño de los experimentos y resultados

4.3.1 Experimento con *Stylo*

Para realizar un análisis estadístico con *Stylo* todo el corpus debe encontrarse almacenado en un mismo directorio. Se procede paso a paso insertando los comandos mencionados en el apartado 3.1.2. Se ofrece aquí el valor que se le debe asignar a cada uno de los argumentos y su propósito.

Comando	Argumento	Valor	Descripción
load.corpus	files	"all"	Carga todos los archivos presentes en el directorio
	corpus.dir	"nombredeldirectorio"	Se especifica solo si los archivos están en un subdirectorio
	encoding	"UTF-8"	La codificación en la que se encuentran todos los archivos
txt.to.words.ext	language	"Spanish"	El corpus está íntegramente en español
	preserve.case	FALSE	Omite la distinción mayúscula / minúscula en el corpus

			transformando todas las mayúsculas en minúsculas
txt.to.features	ngram.size	1	Extrae unigramas, es decir, n-gramas donde n adopta el valor 1
	features	"w"	Extrae n-gramas de palabras, no de caracteres
make.frequency.list	head	5000 / valor numérico igual o superior a 250	Limita a 5000 (valor por defecto de *Stylo*) el número de elementos más frecuentes. En este experimento se emplearán los 250 tokens más frecuentes comenzando por el primero más frecuente, por lo que será indiferente el valor con tal de que sea igual o superior a 250
make.table.of.frequencies	features	Nombre de la variable en la que se haya almacenado el paso anterior	Lista de unigramas más frecuentes a partir de la cual se construirá la tabla de frecuencias
stylo	analysis.type	"CA"	El análisis deseado es el de *cluster* o agrupamiento
	frequencies	Nombre de la variable en la que se haya almacenado el paso anterior	Matriz de frecuencias a partir de la cual se ejecutará la medida Delta
	mfw.max	250	El valor acordado en la configuración previa. El máximo y el mínimo deben coincidir para obtener un solo dendograma
	mfw.min	250	El valor acordado en la configuración previa. El máximo

			y el mínimo deben coincidir para obtener un solo dendograma
	`gui`	FALSE	No aparece la interfaz gráfica
	`custom.graph.title`	"Corpus de prueba"	Título del dendograma
	`write.png.file`	TRUE	Exportar en un fichero de formato .png el dendograma

Tabla 10. Comandos de R con sus respectivos argumentos y descripción de las acciones que ejecutan al introducir esos valores

El comando `txt.to.features` es prescindible puesto que para este análisis se extraen unigramas de palabras, es decir, tokens, por lo que ya en el paso anterior se lleva a cabo esa acción. Empleando la GUI (*graphical user interface*) con los ajustes pertinentes se obtiene de manera mucho más rápida y cómoda el mismo resultado. El dendograma resultante de todo este proceso es este:

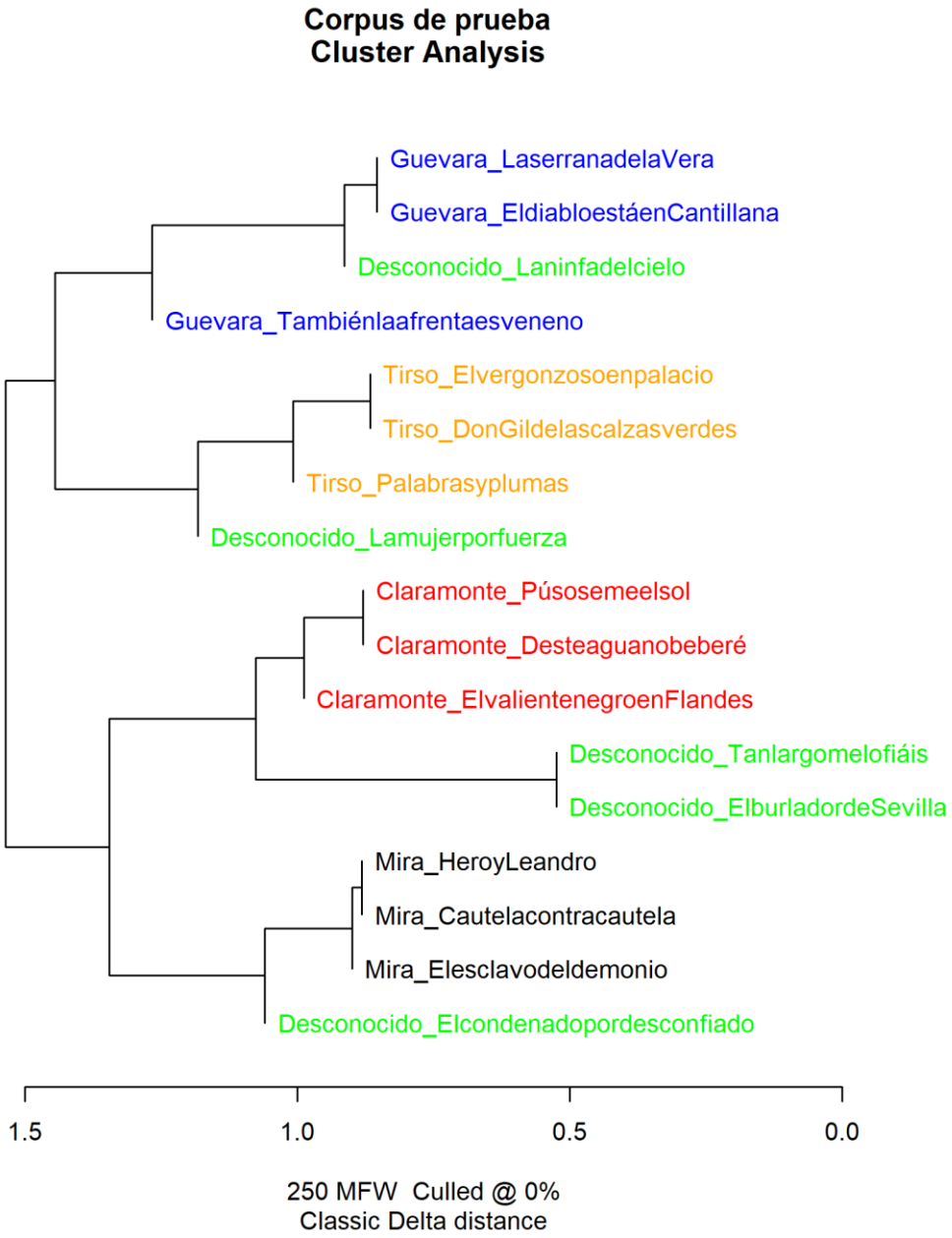


Figura 7. Dendrograma del corpus de prueba con los 250 tokens más frecuentes

Como era esperable, la distancia más pequeña es la que separa *El burlador de Sevilla y Tan largo me lo fiáis*, ya que son, en un porcentaje elevado, el mismo texto. Se agrupan sin lugar a dudas con la producción de Claramonte. Mira se revela como el autor con un estilo más firme en virtud de la escasa distancia entre sus tres obras de autoría asegurada, mientras que Guevara presenta el estilo más inestable. Las parejas Guevara-Tirso y Claramonte-Mira resultan ser las que más similitudes estilísticas comparten. En cuanto a las atribuciones de las obras de autoría desconocida los

resultados obtenidos son rotundos: *El condenado por desconfiado* se asocia con las obras de Mira, *La mujer por fuerza* se aglutina con el corpus tirsiano y *La ninfa del cielo* se une a Guevara, siendo esta última la atribución más categórica de todas.

Como se comentó en el apartado 2, la nueva medida Delta emendada con las implementaciones de Eder (2016) arrojó en algunos experimentos mejores resultados. Para contrastar los obtenidos con la clásica Delta de Burrows conviene llevar a cabo otro análisis similar pero con la de Eder, lo que se puede llevar a cabo de manera ágil gracias a la GUI que ofrece *Stylo*, seleccionando «Eder» en la pestaña *statistics*. Se logra así este dendograma:

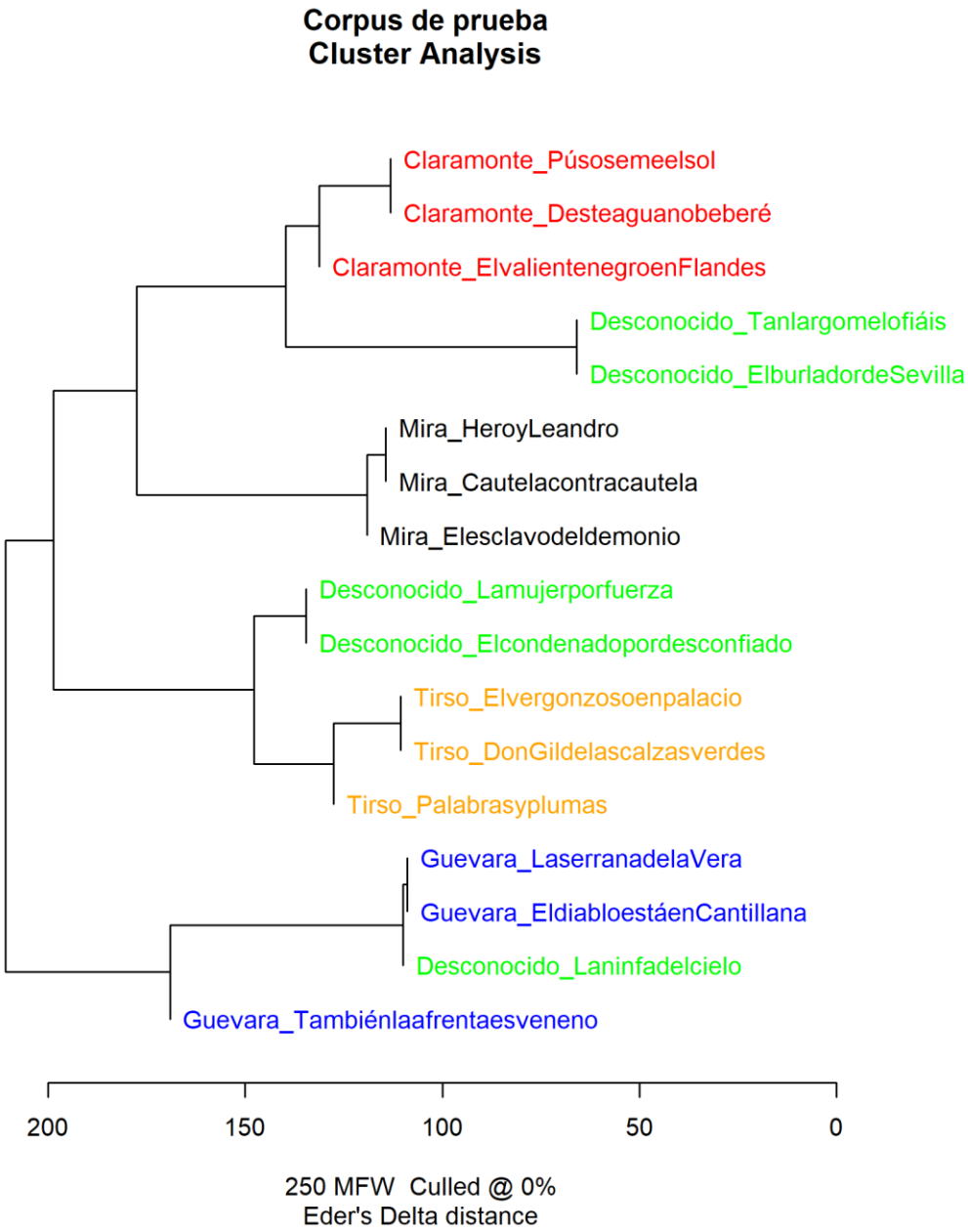


Figura 8. Dendograma del corpus de prueba con los 250 tokens más frecuentes y la distancia Delta Eder

Los cambios con respecto al dendograma anterior son escasos pero notables. La pareja Claramonte-Mira sigue compartiendo un alto grado de similitud, pero no ocurre lo mismo con la formada por Tirso y Guevara, autor que ahora diverge más de los otros tres autores. La principal desemejanza radica en la comedia *El condenado por desconfiado*, que se muestra muy próxima en estilo a *La mujer por fuerza* y se agrupa como esta con las obras de Tirso. Las otras atribuciones siguen manteniendo su rotundidad, en especial *La ninfa del cielo*, que incluso la refuerza.

4.3.2 Experimento con las medidas de distancia

Para acceder al ejecutable que calcula las cuatro medidas (*perplexity*, divergencia Kullback-Leibler, *ranking* y coseno) desde la terminal de Ubuntu se introduce el comando `sh run.sh Nombredeltextoquesequieracomparar.txt`. Para que el cálculo sea correcto, deben almacenarse en un subdirectorio el texto que se quiere comparar y los modelos de cada autor. Estos modelos son archivos creados empalmando todos los textos de autoría conocida del corpus pertenecientes a un mismo autor. De esta manera, habrá que lanzar el ejecutable tantas veces como textos desconocidos haya, en este caso cinco. Cada una de esas veces será necesario borrar del subdirectorio el texto que ya se ha comparado y sustituirlo por el texto que se va a comparar a continuación, empleando para esto los comandos de Ubuntu `rm` (*remove*) y `cp` (*copy*); así como consultar los ficheros de los resultados, que se reescribirán en cada operación. Esos ficheros son cinco, cuatro de las medidas de distancia y uno de la media entre ellas. Todo el proceso puede durar varios minutos. Los resultados que se obtienen, medida por medida, son los siguientes:

	La ninfa del cielo	*El burlador de Sevilla*	*Tan largo me lo fiáis*	*La mujer por fuerza*	*El condenado por desconfiado*
Perplexity	0,0000 Mira	0,0000 Clar	0,0000 Clar	0,0000 Mira	0,0000 Clar
	0,0483 Tirso	0,1525 Tirso	0,5612 Tirso	0,1739 Tirso	0,2502 Mira
	0,4193 Clar	0,6602 Mira	0,7078 Guev	0,7770 Clar	0,6701 Tirso
	1,0000 Guev	1,0000 Guev	1,0000 Mira	1,0000 Guev	1,0000 Guev
Kullback Leibler	0,0000 Mira	0,0000 Clar	0,0000 Clar	0,0000 Tirso	0,0000 Mira
	0,5770 Clar	0,5322 Guev	0,1294 Tirso	0,1626 Mira	0,5719 Clar
	0,6636 Guev	0,6288 Tirso	0,8314 Guev	0,6457 Guev	0,8423 Guev
	1,0000 Tirso	1,0000 Mira	1,0000 Mira	1,0000 Clar	1,0000 Tirso
Ranking	0,0000 Mira	0,0000 Clar	0,0000 Clar	0,0000 Tirso	0,0000 Clar
	0,4000 Guev	0,7895 Guev	0,0238 Mira	0,2149 Clar	0,2410 Mira
	0,8828 Clar	0,8684 Tirso	0,3333 Guev	0,6033 Mira	0,2530 Tirso
	1,0000 Tirso	1,0000 Mira	1,0000 Tirso	1,0000 Guev	1,0000 Guev
Coseno	0,0000 Mira	0,0000 Clar	0,0000 Clar	0,0000 Tirso	0,0000 Mira
	0,3204 Guev	0,0805 Mira	0,4545 Mira	0,0058 Mira	0,1944 Clar
	0,4757 Tirso	0,6356 Tirso	0,6098 Tirso	0,1930 Clar	0,4259 Tirso
	1,0000 Clar	1,0000 Guev	1,0000 Guev	1,0000 Guev	1,0000 Guev

Tabla 11. Resultados normalizados obtenidos con las cuatro medidas de distancia al confrontar las obras de autoría dudosa con los cuatro autores objeto de estudio. Se abrevian los nombres de Claramonte y Guevara para unificar el tamaño de las celdas

Perplexity descarta la autoría de Tirso para todas las obras objeto de estudio, sin embargo lo muestra como segunda opción, a muy poca distancia de la primera, para la mayoría de ellas. La posición del mercedario en las otras medidas tiende a ser menos favorable. Como autor más alejado *perplexity* señala en casi todos los casos a Guevara, incluido curiosamente el de *La ninfa del cielo*, cuya atribución al autor de *El diablo cojuelo* era esperable. Kullback-Leibler concuerda con *Stylo* salvo para *La ninfa del cielo*, que asigna a Mira, igual que hace *ranking* con rotundidad, aunque sitúa a Guevara como la segunda posibilidad más plausible. *Ranking* propone a Claramonte como el autor más posible de *El condenado*, si bien la distancia con Mira y Tirso es pequeña y, entre estos dos, ínfima, en contraste con la diferencia notable que *perplexity* marca entre ambos para la mencionada comedia. No obstante, la principal desemejanza entre

ranking y *perplexity* radica en *La mujer por fuerza*, que *perplexity*, en oposición a las otras medidas adjudica a Mira. Coseno presenta como primera opción exactamente los mismos autores que Kullback-Leibler, pero para las otras opciones tiende a beneficiar más a Mira. Entre *El burlador* y *Tan largo*, las medidas se muestran algo más contundentes para *El burlador*, pero de todos modos rotundas y unánimes en la atribución a Claramonte para ambas versiones. El consenso entre las medidas es notable aunque los valores que arroja *perplexity* divergen ligeramente de los otros. La razón de esta leve disparidad hay que buscarla en los tipos de n-gramas que las estrategias extraen de los textos; mientras que todas las demás emplean unigramas de palabras, *perplexity* utiliza 7-gramas de caracteres (como en el experimento de Gamallo, Pichel y Alegría, 2017). Realizando la media de las cuatro medidas se llega a:

La ninfa del cielo	***El burlador de Sevilla***	***Tan largo me lo fiáis***	***La mujer por fuerza***	***El condenado por desconfiado***
0,000 Mira	0,000 Clar	0,000 Clar	0,043 Tirso	0,123 Mira
0,596 Guev	0,571 Tirso	0,575 Tirso	0,193 Mira	0,192 Clar
0,631 Tirso	0,685 Mira	0,620 Mira	0,546 Clar	0,587 Tirso
0,720 Clar	0,830 Guev	0,718 Guev	0,911 Guev	0,961 Guev

Tabla 12. Media aritmética de los resultados normalizados obtenidos con las cuatro medidas al confrontar las obras de autoría dudosa con los cuatro autores objeto de estudio

La ninfa del cielo, cuya autoría para Guevara está en la teoría casi asegurada, no parece acercarse tanto al estilo del autor de *El diablo cojuelo* sino que presenta más confluencias con Mira de Amescua. Los resultados para *El burlador de Sevilla* y *Tan largo me lo fiáis* son suficientemente categóricos como para desdeñar la supuesta autoría de Tirso de Molina. *La mujer por fuerza* y *El condenado por desconfiado* son las obras que dejan más lugar a dudas e hipótesis alternativas, ya que las cuatro distancias que presentan difieren en una cantidad muy pequeña las unas de las otras y la que separa al autor que alcanza la primera posición del que queda en segundo puesto es menor o igual que 0,15. Mira y Tirso se revelan como los principales contendientes para estas dos comedias. Se puede observar que la divergencia Kullback-Leibler es la medida que arroja resultados más próximos a los medios, mientras que *perplexity* es la que más se aleja. Sin embargo, en el experimento previo con el corpus de entrenamiento

perplexity se revelaba como una estrategia más robusta que Kullback-Leibler. Por tanto, no se pueden extraer conclusiones sobre cuál es la mejor medida para casos de atribuciones de autoría.

La media de las posiciones logradas por los autores es:

La ninfa del cielo	***El burlador de Sevilla***	***Tan largo me lo fiáis***	***La mujer por fuerza***	***El condenado por desconfiado***
1,25 Mira	1 Clar	1 Clar	1,25 Tirso	1,5 Mira
2,75 Guev	2,75 Tirso	2,75 Tirso	2 Mira	1,5 Clar
3 Clar	3 Guev	3 Mira	3 Clar	3,25 Tirso
3,25 Tirso	3,25 Mira	3,25 Guev	3,75 Guev	3,75 Guev

Tabla 13. Media aritmética de las posiciones de los cuatro autores en los resultados obtenidos al confrontarlos con cada obra de autoría desconocida

Se ve así reforzada la atribución de *La ninfa del cielo* y *El condenado por desconfiado* a Mira, así como la de *La mujer por fuerza* a Tirso. *El burlador* y *Tan largo* presentan resultados casi idénticos que apuntan rotundamente a secundar la postura crítica que defiende la autoría de Claramonte.

4. 4 Discusión

Corresponde ahora realizar un cotejo conjunto de los diversos resultados obtenidos por todas las estrategias que permita extraer conclusiones determinantes para dirimir las autorías de los textos. Las autorías propuestas por los diferentes métodos son, simplificando la cuestión:

	Estudios tradicionales	*Stylo*	Medidas de distancia
La ninfa del cielo	Tirso / Guevara	Guevara	Mira
El burlador de Sevilla	Tirso / Claramonte	Claramonte	Claramonte
Tan largo me lo fiáis	Tirso / Claramonte	Claramonte	Claramonte
La mujer por fuerza	Tirso / Otro	Tirso	Tirso
El condenado por desconfiado	Tirso / Claramonte / Guevara / Mira / Colaboración	Mira	Mira

Tabla 14. Atribuciones de autoría de las obras objeto de estudio según las diferentes estrategias

La ninfa del cielo, cuya restitución a Guevara parecía definitiva, no obtiene los resultados esperados con las medidas de distancia, aunque los dendogramas de *Stylo* sí que la agrupan con el resto de obras de Guevara, siendo además la atribución más rotunda de todas. Aunque no se puede excluir la figura de Mira de Amescua, frecuentemente desatendida en los debates críticos del teatro del Siglo de Oro, no parece muy probable que esté detrás de esta comedia. En cambio, los argumentos a favor de Guevara, muy repetidos y contundentes, se ven fortalecidos por este estudio.

El burlador y *Tan largo*, las obras más importantes de la contienda, obtienen resultados muy esclarecedores. La hipótesis defendida insistentemente por Rodríguez López-Vázquez durante tantos años tiene muchas probabilidades de ser cierta. Todas las medidas apuntan de forma rotunda al dramaturgo Andrés de Claramonte, figura ignorada y denostada por la crítica que, en vista de estos resultados, merece al menos una revalorización de su obra y la reconsideración de su posición en el canon de la literatura española. Siendo rigurosos no se debe afirmar, a partir de este estudio, que Claramonte sea efectivamente el autor del primer don Juan, pero sí se puede asegurar que la obra más célebre de Tirso de Molina no fue escrita por Tirso de Molina, ya que los otros autores insertos en el corpus (Guevara y Mira) poseen un estilo tan cercano a la famosa comedia como el del mercedario.

La mujer por fuerza y *El condenado por desconfiado* deben ser consideradas conjuntamente. Como se explicó en el apartado 4.1, una de ellas es la cuarta obra de Tirso en su segunda parte de comedias, por lo que no pueden pertenecer ambas al fraile. En vista de que ningún otro autor ha sido propuesto para *La mujer por fuerza* y de los

resultados que ha obtenido, parece que lo más probable es que esta sea esa cuarta obra que sí pertenece a Tirso. De *El condenado* se han propuesto tantas alternativas que resulta arriesgado apostar por una. Aunque este estudio sitúa a Mira como el candidato más probable y descarta, por la atribución anterior, a Tirso, también hay que considerar como factible la hipótesis de colaboración, que no se puede verificar con las estrategias aquí empleadas.

En definitiva, a Tirso se le han atribuido una cantidad considerable de obras en base a conjeturas y argumentaciones críticas carentes de solidez y demostración documental, por lo que, hoy en día, un elevado porcentaje de la producción que consideramos tirsiana no pertenece en realidad a Tirso de Molina. Si en el siglo XVII se editaban comedias bajo el nombre de autores célebres para incrementar las ventas con plena conciencia del desbarajuste que se estaba cometiendo, la crítica literaria posterior se ha dejado llevar por esa tendencia personalista al favorecer las atribuciones a autores de renombre. Lo curioso del caso de Tirso de Molina es que estas atribuciones falaces se han ido elaborando unas sobre otras, de manera que poner en duda una implica cuestionarlas todas; y que se da la casualidad de que las obras polémicas son precisamente sobre las que más se sustenta su fama entre el público y su excelsa valoración crítica. Por lo tanto, parece necesario, como consecuencia inmediata de este estudio y de otros ya citados que van en la misma dirección, impulsar estas dos acciones:

- La reconsideración del lugar que ocupa Tirso de Molina en la historia de la literatura española
- Una vehemente llamada de atención sobre los estudios de atribución de autoría tradicionales, que en muchas ocasiones no han respetado los principios básicos de rigor científico que deberían regir cualquier clase de investigación humanística

5. CONCLUSIONES Y TRABAJO FUTURO

Una investigación en NTAAS (*non-traditional authorship attribution studies*) debería constar de tres dimensiones:

- Ecdótica: configurar el corpus de estudio y encontrar (o realizar si no están disponibles) las ediciones de los textos más fieles al original
- Computacional: escoger la estrategia que mejor se ajuste al experimento o incluso crearla si fuese necesario
- Experimental: aplicar las estrategias a los textos y analizar los resultados obtenidos teniendo en cuenta el estado de la cuestión del dilema teórico elegido

Para este trabajo, por razones obvias de espacio, tiempo y medios, se ha minimizado el grado de complejidad de las tareas ecdótica y computacional para centrarse en la parte experimental. No obstante, teniendo en cuenta la escasa disponibilidad de ediciones digitalizadas de comedias del siglo de Oro y la fase incipiente en la que se encuentra la disciplina de los NTAAS, especialmente en el ámbito hispánico, la consecución completa de las tareas ecdótica y computacional conllevaría una ardua y prolongada labor. Por eso la forma más evidente de continuar con la investigación sería replicar este experimento a gran escala, de manera que este trabajo sirva como una iniciativa, un estímulo, un punto de partida para desarrollar un amplio estudio que esclarezca al fin los enigmas que rodean al teatro de Tirso. Para ello habría que:

→ Favorecer la edición de textos dramáticos auriseculares. Una fracción considerable del teatro español del siglo de Oro permanece aún sin editar, de modo que el problema no compete únicamente a las ediciones digitales, sino que se extiende también al papel.

→ Expandir el corpus de estudio de este experimento. Con más comedias de los autores seleccionados las conclusiones obtenidas serían más rotundas. En la contienda se podrían incluir más autores hasta conformar un macrocorpus de comedias del siglo XVII. Como mínimo, otras obras dudosas de Tirso que sería clave tratar son *La venganza de Tamar*, *El rey don Pedro* y *Siempre ayuda la verdad*.

→ Buscar configuraciones alternativas de las estrategias que sean eficaces e incluso crear nuevas estrategias más adecuadas para el estudio de atribuciones de

autoría. Si bien *Stylo* se revela como una herramienta potente que fue concebida específicamente para experimentos de este tipo, sus limitaciones todavía son notables. Potenciar la investigación interdisciplinar en NTAAS, de la mano de la lingüística computacional y la lingüística forense, fomentaría que la disciplina se consolidase definitivamente.

→ Reforzar el consenso entre los investigadores que llevan años ocupándose de este asunto. Revisar atentamente las contribuciones ajenas, discutirlas con argumentos sólidos y demostrables, no ejercer una ofensiva constante hacia el sector que defiende la postura contraria y no obcecarse en la propia son pequeñas acciones que favorecen un ambiente de investigación propicio a la aparición de respuestas. Desgraciadamente, el clima que rodea al debate tirsiano está lejos de seguir estas pautas.

Un estudio a parte merecería la comedia *El condenado por desconfiado*, cuyas hipótesis de autoría son demasiado variadas como para tener cabida en un estudio tan sintetizado como este. Con herramientas que estudiasen trozos de textos se podría determinar si es una obra compuesta en colaboración o si uno de los autores propuestos es efectivamente el auténtico.

Del mismo modo, debería recibir especial atención por parte de los especialistas en literatura española del Siglo de Oro la figura de Andrés de Claramonte. La demostración terminante de que fue el dramaturgo murciano el autor del primer don Juan parece estar más cerca que nunca. Reivindicar su obra, empezando por editarla, sería un acto de justicia crítica. Como corolario a este trabajo se podrían emplear las mismas herramientas para cotejar la autoría de la comedia *La estrella de Sevilla*, tradicionalmente atribuida a Lope pero que muy probablemente pertenezca también a Claramonte. Sería este un primer paso en favor de la reivindicación del autor.

En el terreno puramente computacional, proseguir con la aplicación de las medidas de distancia aquí empleadas —cuya funcionalidad original se adscribe a otros campos de la lingüística computacional— en problemas de atribución de autoría conducirá a su perfeccionamiento y a una valoración razonada de su eficiencia, de manera que se pueda repetir este experimento ponderando las medidas en función de su efectividad y otorgando más valor, por consiguiente, a aquellas que se revelen más aptas, mejorando así la precisión de los resultados obtenidos.

A lo largo del trabajo se ha presentado el paradigma de los NTAAS y el funcionamiento de una serie de estrategias de atribución automática de autoría que han demostrado su rentabilidad en investigaciones lingüísticas y literarias. A partir de uno de los dilemas de autoría más interesantes de la historia de la literatura española, la producción dramática tirsiana, se ha expuesto en primer lugar la polémica teórica que lo rodea para pasar después a la aplicación práctica de las estrategias antes explicadas. De esta manera se combina el estudio tradicional con el no tradicional para que los resultados obtenidos puedan ser interpretados bajo la perspectiva adecuada. Para dicha aplicación se ha configurado un corpus de comedias auriseculares de los cuatro autores inmersos en la polémica y cinco obras de autoría dudosa. Una vez llevada a cabo la experimentación se han presentado los resultados obtenidos de la manera más detallada posible, procurando en todo momento una visualización transparente de los mismos. Finalmente se han cotejado conjuntamente todos los resultados con el fin de extraer conclusiones coherentes que fomenten un debate razonado acerca del problema en cuestión y se ha trazado un plan de futuro para su resolución.

La principal aportación de este trabajo es la reorientación del problema de las falsas atribuciones a Tirso hacia el ámbito de las humanidades digitales, lo que ha permitido arrojar luz sobre las hipótesis que la crítica tradicional lleva siglos proponiendo. Sin posicionarse de antemano por ninguna, los resultados obligan a hacerlo ahora en favor de aquellas teorías que abogan por autores menos célebres que han sido relegados a un segundo plano en el panorama del teatro aurisecular. Entre ellos, Mira de Amescua, Vélez de Guevara y, especialmente, Andrés de Claramonte.

En definitiva, sirva este trabajo como prueba de que la cooperación entre diversas disciplinas produce frutos de provecho para ambas. El desarrollo de la lingüística computacional ha abierto el camino hacia una revolución de los métodos de investigación en humanidades en la era digital y a su vez, el esclarecimiento de dilemas autorales aportará datos valiosos para la historiografía literaria, el análisis histórico del discurso o la misma lingüística computacional, madurando estrategias que podrán ser reaprovechadas en investigaciones que persigan otros propósitos.

BIBLIOGRAFÍA

ALMEIDA, Dayane Celestino de (2014). «Atribuição de autoria com propósitos forenses: panorama e proposta de análise». *ReVEL: Revista Virtual de Estudos da Linguagem*, [en línea], 12, 23, pp. 148-186, <http://www.revel.inf.br/files/539b2f0878d56cb6604363c111dfe116.pdf>, [Consulta: 11/04/19]

BLASCO, Javier (2016). «Avellaneda desde la estilometría». En: Pedro Ruiz (ed.), *Cervantes: los viajes y los días*. Madrid: Sial Ediciones, pp. 97-116.

BLOOMFIELD, Leonard (1984)[1933]. *Language*. Chicago: University of Chicago Press.

CALVO TELLO, José (2016). «Entendiendo Delta desde las Humanidades». *Caracteres. Estudios culturales y críticos de la esfera digital*, [en línea], 5, 1, pp. 140-176, <http://revistacaracteres.net/revista/vol5n1mayo2016/entendiendo-delta/>, [Consulta: 04/04/19]

EDER, Maciej, Jan Rybicki y Mike Kestemont (2016). «Stylometry with R: a package for computational text analysis». *R Journal*, [en línea], 8, 1, pp. 107-121, <https://journal.r-project.org/archive/2016/RJ-2016-007/index.html>, [Consulta: 04/04/19]

GAMALLO, Pablo, José Ramom Pichel e Iñaki Alegría (2017). «From language identification to language distance». *Physica A*, [en línea], 484, pp. 152-162, <https://www.sciencedirect.com/science/article/pii/S0378437117305137>, [Consulta: 04/04/19]

GAMALLO, Pablo, Marcos García, Susana Sotelo y José Ramom Pichel (2014). «Comparing Ranking-based and Naive Bayes Approaches to Language Detection on Tweets». En: Arkatiz Zubiaga, Iñaki San Vicente et. al. (eds.), *Actas del XXX Congreso de la Sociedad Española de Procesamiento de lenguaje natural*, [en línea]. CEUR Workshop Proceedings, 1228, pp. 12-16, <http://ceur-ws.org/Vol-1228/tweetlid-1-gamallo.pdf>, [Consulta: 13/06/19]

GARCÍA-REIDY, Alejandro (2019). «Deconstructing the Authorship of *Siempre ayuda la verdad*: A Play by Lope de Vega?». *Neophilologus*, [en línea], s.v., pp. 1-18, <

https://link.springer.com/content/pdf/10.1007%2Fs11061-019-09607-8.pdf> [Consulta: 04/06/19]

GARCÍA GÓMEZ, Ángel María (2005). «Aporte documental al debate acerca de la prioridad entre *El burlador de Sevilla* y *Tan largo me lo fiáis*: el cartapacio de comedias de Jerónimo Sánchez». En: Anthony J. Close & Sandra María Fernández Vales (eds.), *Edad de oro cantabrigense: actas del VII Congreso de la Asociación Internacional de Hispanistas del Siglo de Oro*. Madrid: Asociación Internacional del Siglo de Oro (AISO), pp. 281-286.

GRIEVE, Jack (2005). *Quantitative Authorship Attribution: a History and an Evaluation of Techniques* [en línea]. Burnaby: Simon Fraser University Institutional Repository, <http://summit.sfu.ca/item/8840>, [Consulta: 04/04/19]

IRIARTE, Álvaro, Pablo Gamallo y Alberto Simões (2018). «Estratégias Lexicométricas para Detetar Especificidades Textuais». *Linguamática*, [en línea], 10, 1, pp. 19-26, <https://linguamatica.com/index.php/linguamatica/article/view/263>, [Consulta: 04/04/19]

LA ROSA, Javier de y Juan Luis Suárez (2016). «The Life of *Lazarillo de Tormes* and of His Machine Learning Adversities: Non-traditional authorship attribution techniques in the context of the *Lazarillo*». *Lemir: Revista de Literatura Española Medieval y del Renacimiento*, [en línea], 20, pp. 373-438, <http://www.cultureplex.ca/wp-content/uploads/2016/09/09_Rosa_Javier_de_la.pdf >, [Consulta: 06/04/19]

LOVE, Harold (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

MANNING, Christopher, Prabhakar Raghavan y Hinrich Schütze (2008). *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

MOLINA, Tirso de y Luis Vélez de Guevara (2008)[1610-1635]. *El condenado por desconfiado – La ninfa del cielo*, ed. Alfredo Rodríguez López-Vázquez. Madrid: Cátedra.

OTEIZA, Blanca (2000). «¿Conocemos los textos verdaderos de Tirso de Molina?». En: Ignacio Arellano & Blanca Oteiza (eds.), *Varia lección de Tirso de Molina (Actas del*

VIII Seminario del Centro para la Edición de Clásicos Españoles). Pamplona: Instituto de Estudios Tirsianos, pp. 99-128.

RODRÍGUEZ LÓPEZ-VÁZQUEZ, Alfredo (1983). «La autoría de *El burlador de Sevilla*: Andrés de Claramonte». *Castilla: Estudios de Literatura*, 5, pp. 87-108.

— (1990). «El estado de la cuestión en torno a Clarmonte y *El burlador de Sevilla*». *Murgetana*, [en línea], 82, pp. 5-22, <https://www.academia.edu/22111004/EL_ESTADO_DE_LA_CUESTI%C3%93N_EN_TORNO_A_CLARAMONTE_Y_EL_BURLADOR_DE_SEVILLA_POR_ALFREDO_RODR%C3%8DGUEZ_L%C3%93PEZ-V%C3%81ZQUEZ>, [Consulta: 24/03/19]

— (2010). «*La mujer por fuerza, El condenado por desconfiado* y *El burlador de Sevilla*, tres comedias atribuidas a Tirso de Molina». *Castilla. Estudios de Literatura*, [en línea], 1, pp. 131-153, < https://www.academia.edu/8666821/La_mujer_por_fuerza_El_condenado_por_desconfiado_y_El_burlador_de_Sevilla_tres_comedias_atribuidas_a_Tirso>, [Consulta: 24/03/19]

— (s. f.). «*El burlador de Sevilla, La estrella de Sevilla* y los problemas de su autoría», s. r., [en línea], s. n., pp. 9-25, <http://www.academia.edu/10229995/El_Burlador_de_Sevilla_la_Estrella_de_Sevilla_y_los_problemas_de_su_autor%C3%ADa>, [Consulta: 24/03/19]

RUANO DE LA HAZA, José María (1995). «La relación textual entre *El burlador de Sevilla* y *Tan largo me lo fiáis*». En: Ignacio Arellano, Blanca Oteiza, Carmen Pinillos & Miguel Zugasti (eds.), *Tirso de Molina, del siglo de oro al siglo XX: actas del coloquio internacional: Pamplona, Universidad de Navarra, 15-17 diciembre 1994*. Pamplona: Instituto de Estudios Trisianos, pp. 283-296.

RUDMAN, Joseph (1998). «The State of Authorship Attribution Studies: Some Problems and Solutions». *Computers and the Humanities*, 31, 4, pp. 351-365.

— (2016). «Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats». *Journal of Early Modern Studies*, [en línea], 5, pp. 307-328,

<http://www.fupress.net/index.php/bsfm-jems/article/view/18094/16848>, [Consulta: 04/04/19]

STAMATOS, Efstathios (2009). «A Survey of Modern Authorship Attribution Methods». *Journal of the American Society for Information Science and Technology*, [en línea], 60, 3, pp. 538-556, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.1634&rep=rep1&type=pdf>, [Consulta: 04/04/19]

VÁZQUEZ, Luis (1995). «*El burlador de Sevilla*: claramente de Tirso y no de Claramonte (breve anotación crítica)». *Bulletin of the comediantes*, 47, 2, pp. 183-190.

Las obras extraídas del portal de la Biblioteca Virtual Miguel de Cervantes <http://www.cervantesvirtual.com/> son ediciones digitalizadas que cotejan las ediciones críticas más relevantes de la obra en cuestión. Las obras extraídas de la web de la Association for Hispanic Classical Theater <http://www.wordpress.comedias.org/play-texts/> son ediciones del profesor Vern Williamsen a partir de ediciones antiguas o manuscritos (la procedencia específica se explica en nota al inicio de cada comedia). En la página de la asociación se explicitan los criterios seguidos. Las dos obras que no fueron obtenidas de ninguno de estos repertorios por no estar disponibles son:

CLARAMONTE, Andrés de (2011)[1638]. *El valiente negro en Flandes*, ed. Enrique Escudero & Nancho Novo, [en línea]. Bubok Publishing, <https://www.bubok.es/libros/199348/El-valiente-negro-en-Flandes>, [Consulta: 03/04/19]

MOLINA, Tirso de (2013)[1635]. *La mujer por fuerza*, ed. Elena Garcés Molina, [en línea]. Clásicos Hispánicos, <http://www.clasicoshispanicos.com/tirso-de-molina/68-ebook-tirso-molina-mujer-fuerza.html>, [Consulta: 03/04/19]

En cuanto al software, el paquete *Stylo* de R está disponible para descarga en este enlace: <https://CRAN.R-project.org/package=stylo>, [Consulta: 10/06/19]

Las medidas de distancia *perplexity* y Kullback-Leibler empleadas en los experimentos de Gamallo, Pichel y Alegría (2017) e Iriarte, Gamallo y Simões (2018) respectivamente, están disponibles en la plataforma de desarrollo colaborativo GitHub, en los enlaces: <https://github.com/gamallo/Perplexity>, <https://github.com/ambs/Math-KullbackLeibler-Discrete>, [Consulta: 10/06/19]. La implementación de las cuatro medidas diseñada específicamente para este trabajo se aloja en: <https://github.com/gamallo/Autoria>, [Consulta: 13/06/19]

APÉNDICE

Se exponen a continuación, como curiosidad y para realizar comparaciones a simple vista, los 250 tokens más frecuentes del corpus de entrenamiento (el constituido por las obras de autoría conocida) y los 100 de los respectivos corpus de cada autor objeto de estudio. Para obtener las listas se ha empleado el comando `make.frequency.list` de *Stylo*, aunque al construir dendogramas con la GUI también se almacenan automáticamente en un archivo .txt los ítems (MFW) empleados. El orden de frecuencia, de mayor a menor, sigue el patrón de lectura de izquierda a derecha, de arriba abajo.

que	de	y	el	la
a	en	no	es	me
con	mi	don	por	si
los	qué	un	se	su
yo	rey	lo	ha	las
al	te	pues	juan	del
ya	más	le	amor	tu
doña	para	una	como	porque
tan	señor	os	he	bien
ser	dios	esta	aquí	está
sin	quien	soy	dos	mí
hay	mis	son	este	mas
mujer	quién	él	d	conde
sus	vos	aunque	o	vida
quiero	juana	aparte	alma	cuando
sale	duque	gil	esto	estoy
cielo	ni	pero	hombre	leonor
ella	tú	así	gila	vuestra
mal	muerte	enrico	catalinón	cómo
enrique	tengo	has	casa	padre
todo	tiene	tus	tal	sol
fue	sí	diego	nos	hoy
han	inés	vase	ver	entre
esperanza	ninfa	todos	ay	fin
señora	ti	hacer	sido	noche
ojos	también	teodora	leandro	mar
tanto	salen	sé	finea	luego
gran	vuestro	marqués	honor	carlos
valor	donde	mano	eso	agora
mil	siempre	muy	pedro	da
puede	vanse	roberto	suerte	íñigo
capitán	celos	hero	lope	gallardo
otra	esa	nombre	otro	voy
día	dar	gusto	eres	pienso
ocasión	fuera	viene	vive	escena

madalena	mía	ese	desde	césar
hecho	lisarda	mundo	amigo	hasta
será	sólo	mayor	martín	dado
octavio	pecho	así	gutierre	sois
fuego	garcía	oh	poco	mira
mismo	paulo	verdad	agua	vez
razón	florela	mucho	tiempo	mireno
fe	gente	rodrigo	pedrisco	mencía
dice	causa	matilde	criado	pies
cosa	príncipe	va	cielos	di
mejor	parece	sea	sirena	digo
lugar	espada	hace	muerto	quiere
visto	después	estás	favor	puedo
dónde	nunca	habéis	sangre	alteza
vamos	estos	fernando	loco	quieres

Tabla 1. 250 MFW del corpus de entrenamiento

que	y	de	el	en
la	a	es	no	me
con	los	mi	si	por
don	se	juan	un	rey
su	las	teodora	del	ha
qué	yo	le	pues	te
lo	al	ya	más	gutierre
doña	mencía	tan	señor	aquí
tu	mis	porque	mas	quién
he	agustín	amor	ser	una
lesbia	natalio	mujer	zurdo	leonor
dios	soy	está	mí	sin
así	esta	para	duque	sus
como	son	diego	cuando	él
este	alcina	sale	fidelfo	quiero
sol	aunque	os	bien	dos
cielo	antón	negro	honor	alma
tal	estoy	hay	juana	esto
abad	capitán	ella	vos	fernando
ay	padre	vida	gil	vase

Tabla 2. 100 MFW del corpus de Claramonte

que	de	y	la	el
a	en	no	con	es
rey	por	me	gila	los
mi	las	un	al	don
lo	lope	se	si	qué

esperanza	ha	su	yo	más
del	tan	como	te	pues
esta	para	juan	leonor	rodrigo
una	ya	os	bien	mingo
porque	hay	capitán	dios	garcía
ser	le	vos	vuestra	sin
señor	aquí	he	amor	giraldo
perafán	está	esto	quien	son
tu	este	han	mí	soy
mujer	muy	alteza	aunque	andrés
fernando	serrana	dos	él	vuestro
maría	valor	ella	mis	ni
o	vida	cielo	doña	madalena
hombre	cuando	padre	d	quién
sus	todo	vera	todos	alma

Tabla 3. 100 MFW del corpus de Guevara

que	y	de	la	el
a	no	en	es	me
mi	con	don	los	si
un	te	qué	enrique	ha
se	las	por	yo	amor
leandro	tu	rey	su	lo
del	al	hero	pues	ya
más	porque	lisarda	una	ser
he	le	son	está	señor
dios	gil	para	aparte	bien
como	dos	leonor	diego	césar
esta	floro	sin	marcelo	ludovico
vida	quien	este	soy	alma
tan	mis	sancho	elena	muerte
príncipe	chirimía	domingo	ni	hay
mí	sus	tú	tus	mitilene
o	aquí	él	porcia	quién
pero	quiero	mas	tiene	cielo
has	os	así	vase	escena
mal	fue	nos	polidoro	cuando

Tabla 4. 100 MFW del corpus de Mira

que	de	y	a	el
en	la	no	don	mi
es	doña	me	por	con
si	su	qué	un	juana
yo	lo	amor	los	pues

os	ha	se	al	inés
las	ya	te	le	íñigo
gil	más	para	del	como
duque	rey	tu	gallardo	aparte
bien	quien	mas	mireno	martín
he	conde	matilde	sin	aquí
una	vos	ser	porque	él
sirena	dos	mí	esta	hay
madalena	o	señor	caramanchel	tan
antonio	dios	juan	mis	tarso
está	casa	aunque	pero	serafina
este	ni	vuestra	laura	soy
esto	sus	quintana	vuestro	sí
vida	alma	fin	quiero	ella
pedro	cuando	son	hombre	próspero

Tabla 5. 100 MFW del corpus de Tirso